Table of Contents

Foreword

Jonathan Webster, City University of Hong Kong
Nancy Ide, Vassar College
Alex Chengyu Fang, City University of Hong Kong

Language resources, including not only corpora but also lexicons, knowledge bases and ontologies, grammars, etc. support the development of language processing applications that are increasingly important to the global society. Substantial effort has been devoted to the creation of such resources for the world's major languages over the past decades, and new projects are developing similar resources for less widely-used languages. Some standards and best practices have emerged for representing and linking language corpora and annotations, efforts such as "Global WordNet" and the development of framenets in multiple languages seek to create and link specific lexical and semantic resources across languages, and there are efforts to integrate such resources into general ontologies such as SUMO and to enable common access to ontologies spread across the World Wide Web. As the need for cross-lingual studies and applications grows, it is increasingly important to develop resources in the world's languages that can be compared and linked, used and analyzed with common software, and that contain linguistic information for the same or comparable phenomena. We envision the eventual development of a "global web" of language resources, wherein, for example, linguistically-annotated corpora in multiple languages are inter-linked via the use of common categories, or categories that are mapped to one another; resources such as wordnets and framenets are linked not only to versions in different languages, but also to each other; and common representations enable analysis and use of resources in different languages and of different types within available systems.

The First International Conference on Global Interoperability for Language Resources will bring together designers, developers, and users of corpora and other language resources from across the globe, in order to:

- assess the state of the art in methods and schemes for resource representation, annotation, interlinkage, and access;
- consider the requirements for (and obstacles to) multi-lingual and multi-modal interoperability and standardization;
- consider the requirements for achieving interoperability among multi-lingual resources of different types, including corpora, lexicons, knowledge bases, ontologies, etc., as well as the systems and frameworks that enable their creation and exploitation;
- consider the ways in which web technologies are and may be used to enable resource interoperability and inter-linkage;

- work toward the definition of best practice guidelines and standards for language resource representation, annotation, and use that will enable interoperability;
- consider means to map or harmonize linguistic information in order to better enable cross-lingual studies;
- provide direction for developers of resources for less widely used languages;
- promote collaboration and cooperation among developers of language resources across the globe;
- consider ways to provide central or distributed access to language resources developed throughout the world.

Topics

Paper submissions are invited on (but not limited to) the following topics:

- multi-lingual and/or multi-modal language resources, with focus on the mechanisms enabling interoperability;
- support for multi-linguality and multi-modality in systems/frameworks for resource creation, annotation, use, and access;
- existing and proposed standards and best practice guidelines for language resources, including standards for linguistic annotations at any and all
- linguistics levels;
- systems, frameworks, and architectures to support the development and exploitation of interoperable language resources;
- evaluation of existing resources, systems and frameworks, and/or standards in terms of support for interoperability;
- harmonization, integration, and/or linking of language resources, including corpora, wordnets, framenets, ontologies, etc.;
- web-based technologies for resource interoperability, inter-linkage, and access;
- interoperability of ontologies for language processing research;
- support for multi-linguality, multi-culturality, and multi-modality.

In addition to full-length paper presentations, the Program Committee also invites proposals for posters addressing any of the above topics. Posters describing existing or developing resources or tools that provide an assessment of needs and/or considerations for interoperability are especially encouraged.

Acknowledgment

The ICGL Conference Organisers would like to acknowledge the generous support received from:

City University of Hong Kong
Hong Kong SAR

Department of Chinese, Translation and Linguistics
City University of Hong Kong
Hong Kong SAR

The Halliday Centre of Intelligent Applications for Language Studies
City University of Hong Kong
Hong Kong SAR

Vassar College
USA

China National Institute of Standardization
China

Invited Speakers

Collin Baker
International Computer Science Institute, UC Berkeley
USA
collinb@ICSI.Berkeley.edu

Nicoletta Calzolari
Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche
Italy
glottolo@ilc.cnr.it

Christiane Fellbaum
Princeton University
USA
Fellbaum@princeton.edu

Nancy Ide
Vassar College
USA
ide@cs.vassar.edu

D. Terence Langendoen
University of Arizona
USA
dlangend@nsf.gov

Eric Nyberg
Carnegie Mellon University
USA
ehn@cs.cmu.edu

Programme Committee

Eric Atwell, Leeds University, UK
Harry Bunt, the University of Tilburg, the Netherlands
Bran Bogureav, IBM, USA
Nicoletta Calzolari, Consiglio Nazionale delle Ricerche, Italy
Key-Sun Choi, Korea Advanced Institute of Science and Technology, South Korea
Khalid Choukri, Evaluations and Language Resources Distribution Agency, France
Chris Cieri, Linguistic Data Consortium, USA
Arienne Dwyer, University of Kansas, USA
Alex Chengyu Fang, City University of Hong Kong, Hong Kong SAR
Christiane Fellbaum, Princeton University, USA
Charles Fillmore, International Computer Science Institute, UC Berkeley, USA
Sadaoki Furui, Tokyo Institute of Technology, Japan
Eva Hajicova, Charles University, Czech Republic
Erhard Hinrichs, Eberhard Karls Universität Tübingen, Germany
Mark Huckvale, University College London, UK
Nancy Ide, Vassar College, USA
Hitoshi Isahara, National Institute of Information and Communications Technology, Japan
Toru Ishida, Kyoto University, Japan
Kiyong Lee, Korea University, South Korea
Duo Li, Peking University, China
Inderjeet Mani, Georgetown University, USA
Srini Narayanan, International Computer Science Institute, UC Berkeley, USA
Adam Pease, Articulate Software, USA
Sameer Pradhan, BBN Technologies, USA
James Pustejovsky, Brandeis University, USA
Laurent Romary, Max-Planck Digital Library, Germany
Vasile Rus, the University of Memphis, USA
Pavel Smrz, Brno University of Technology, Czech Republic
Maosong Sun, Tsinghua University, China
Takenobu Tokunaga, Tokyo Institute of Technology, Japan
Piek Vossen, Vrije University, the Netherlands
Jonathan Webster, City University of Hong Kong, Hong Kong SAR
Peter Wittenburg, Max-Planck Institute for Psycholinguistics, the Netherlands
Yihua Zhang, Guangdong University of Foreign Studies, China

Conference Organisers

Conference Convener: **Jonathan Webster**
City University of Hong Kong
Hong Kong SAR
ctjjw@cityu.edu.hk

Conference Co-chairs: **Nancy Ide**
Vassar College
USA
ide@cs.vassar.edu

Alex Chengyu Fang
City University of Hong Kong
Hong Kong SAR
acfang@cityu.edu.hk

Conference Secretary: **Ernest Shue Fung Lam**
City University of Hong Kong
Hong Kong SAR
ernestl@cityu.edu.hk

Conference Webmaster: **Kin Tat Ko**
City University of Hong Kong
Hong Kong SAR
ctktko@cityu.edu.hk

Conference Designer: **Chris Pak-hang Leung**
City University of Hong Kong
Hong Kong SAR
phleung@cityu.edu.hk

Papers

FrameNet, Present and Future

Collin Baker
FrameNet Project
International Computer Science Institute
Berkeley, California
collinb@icsi.berkeley.edu

Abstract

This paper will focus on recent and near-term future developments at FrameNet (FN) and the interoperability issues they raise. We begin by discussing the current state of the Berkeley FN database, the data formats and APIs available, and the relations between FN grammatical functions and "standard" parses and between FN frame elements and "standard" semantic/thematic/theta roles.

We also cover five projects currently underway at ICSI and the interoperability/ compatibility issues connected with them: (1) "Rapid Vanguarding", which aims to integrate Adam Kilgarriff's (2002, 2003) Word Sketch Engine into the frame and lexical unit definition process (2) "Beyond the Core", which is developing tools for annotating constructions, especially those which are neither simply lexical nor easy for the standard parsers to parse, (3) FN's part in the development of the Manually Annotated Subcorpus of the American National Corpus, (4) a pilot study on aligning WordNet and FrameNet, to exploit the complementary strengths of these quite different resources, and (5) a study of image-schematic frames and their usefulness in inferencing.

We discuss FN-related research on Spanish, Japanese, German (SALSA) and other languages, and the putative language-independence of frames, along with other interesting FN-related work by others, and a sketch of a large group of image-schematic frames which are now being added to FN. We close with some ideas about how FrameNet can be opened up, to allow broader participation in the development process without losing precision and coherence.

1 The FrameNet Database

FrameNet (hereafter FN) is a lexicon of English which is intended to be both human- and machine-readable, based on the theory of frame semantics (Fillmore, 1982), which asserts that the meanings of many words are best understood in terms of an entire situation and the participants and props involved in it; the situation is called a **frame**, and the participant roles are called **frame elements (FEs)**. The link between a lemma and a frame is a **lexical unit (LU)**, which is roughly equivalent to a word sense in a conventional dictionary, or to a WN sense (although these three types of resource are designed on different principles and so make different choices about dividing senses).

Instances of the lexical units are manually annotated, marking target word and the occurrence (or in some cases, non-occurrence) of frame elements in each sentence. Then reports are generated showing the possible valences of each lexical unit.

The FrameNet database, as of Nov. 13, 2007, contained 798 lexical semantic frames, covering 11,141 lexical units, or roughly 14 lexical units per frame. The frames are linked to each other with a variety of frame relations, including several types of inheritance, and a further 86 non-lexical frames have

been created to fill out the frame hierarchy. Altogether, these 884 frames contain 8,258 frame elements (frame-specific roles), or about 9 per frame.

There are currently 166,270 annotated instances of lexical units (LUs) in the database. Roughly 85% of these are "lexicographic" annotation, in which only one LU is annotated per sentence, and roughly 10-20 sentences have been annotated for each LU, selected so as to show the full range of valences for the LU. The other 15% of the instances are running text, which is annotated for all the LUs in each sentence, called "full-text" annotation. The sentences for the lexicographic annotation are drawn largely from the British National Corpus; those for full-text annotation are drawn from several sources, including the American National Corpus, the Nuclear Threat Initiative website (www.nti.org), and the Wall Street Journal. The database is stored in MySQL, with annotation carried out using a Java GUI client connecting via an application server implemented in JBOSS.

1.1 Importing text

The importation of sentences into the database involves first converting the raw text into XML, then inserting it into the appropriate tables. Since the requirements for lexicographic annotation and full-text annotation are different, these processes involve different steps, although some of the tools used are the same.

There are several low-level encoding issues: We began work with an early version of the BNC and continue to use it for lexicographic work, which results in some problems with incompatibility that we have not yet solved. The original encoding of the BNC was SGML, rather than XML, and ISO-8859-1 rather than Unicode. When retrieving example sentences for lexicographic annotation, we retrieve them with the BNC (CLAWS) POS tags and in ISO-8859-1 encoding. This is causing problems as we attempt to convert everything to Unicode.

The full texts come from a variety of sources. Those from the ANC are nominally Unicode in UTF-16; we convert these to UTF-8 for importing, since our MySQL database and several other pieces of software are set up for UTF-8. There are, however, a small proportion of non-Unicode characters in some of the ANC texts; so far there have been so few of these that we have been correcting them by hand, but we hope to find a more reliable way of dealing with them automatically.

1.2 Reports and Data Releases

The FrameNet software can produce various types of reports; some are for use internally or on the FN website; others constitute the main data distribution formats:

Single files	
frames.xml	complete definitions of all frames and FEs
frRelation.xml	gives all relations between frames: inheritance, causative of, etc.
semtypes.xml	a small hierarchy of semantic types that can be applied to frames, FEs, or LUs.
Per document	
corpus.xml	one file for each document for full-text annotation
Per lexical unit	
luNnn.xml	XML of the text of the sentence and all layers of annotation for the lexicographic examples.
luNnnPOS.xml	same as luNnnxml, but also has POS tags, twice as big.

Most of the files listed above have a corresponding HTML file, so that a browsable version of the FN database, comparable to the public website, can be downloaded and set up locally by any user. There is also a set of DTDs for the above XML file formats and rather extensive documentation. What is urgently needed is a set of APIs for various programmins languages, so that users will no longer have to build their own.

1.3 FN frame elements and "standard" semantic/thematic/theta roles

FN data users and visitors to our website often ask "What is the relation between the 8,000+ FEs in the FrameNet database and the eight or ten case roles of Fillmore's early work (Fillmore, 1968)?" The real answer is that almost all of the FEs are connected through a series of FE-to-FE links that go along with

the frame-to-frame hierarchy to a high-level frame, such as Event, Action, Intentionally_act, Motion, etc. The FEs in these high-level frames are named Agent, Theme, Source, Path, Goal, Manner, Means, Instrument, etc. thus covering roughly the basic case roles. However, this requires traversing the links to find out what case role a given FE belongs to–and there are some FEs that do **not** link to high-level frames for all the FEs, such as the Similarity frame, home to LUs such as *like*.a and *resemble*.v. Similarity inherits from the frames Gradable_attributes and Reciprocality, but neither of those will supply anything like the traditional case roles, as they are simply not applicable to the Resemblance frame.

2 Current projects at ICSI

2.1 'Rapid vanguarding"

The FrameNet team is currently engaged in NSF-funded research (IIS-0535297 "Rapid Development of a Frame-semantic Lexicon") to build new software for the "vanguarding" portion of its work, that is, the process of defining new frames and their frame elements and determining what lexical units they are evoked by. This can be a very time-consuming process, involving repeated searches of the corpus for each lexical unit. The new tools, modeled on the Word Sketch Engine developed by Adam Kilgarriff and associates (Kilgarriff et al., July 2004), will eliminate duplication of effort and allow decisions about an entire group of homonyms or polysemous words to be made simultaneously.

However the Word Sketch Engine uses patterns of POS tags to extract the information it needs; that means that we need to have consistent POS tagging on our examples, but unfortunately, as they come from different sources, they do not currently have consistent POS labels.

2.2 Syntactic constructions "Beyond the Core"

FN is also beginning an exploratory project on annotating non-lexical constructions. According to the theory of construction grammar and frame semantics, there is only one kind of linguistic object that constitutes what speakers of a language have to learn: the construction, a sign (i.e. a pairing of a form and a meaning). Lexical units are simply constructions whose form pole is one or more word forms, and whose meaning pole is represented as a quite specific semantic frame. In the case of other, non-lexical constructions, such as predication and the genitive, the syntax of the form side is clear, but the meaning of the frame evoked is extremely vague.

Many of the "interesting" constructions have some lexical elements, and they are precisely what cause conventional parsers to fail or give incomplete analyses of sentences such as the following:

(1) a. I can't stand to see, let alone touch, boa constrictors. (*Let alone* functions as a conjunction, but with very specific semantic constraints on the pieces that it joins; this is combined with a Right-node Raising construction. (Fillmore et al., 1988))

b. The gifted have a duty to help the less fortunate. (*The* + Adjective forms a noun meaning "people who have this quality".)

c. What's this scratch doing on the table top? (The scratch isn't **doing** anything, and the construction as a whole carries an implication that there's something odd or wrong about the situation. (Kay and Fillmore, 1999))

We have received funding[1] for a pilot project to document non-lexical constructions just as the current FrameNet documents the lexical constructions, by manually annotating examples drawn from corpora, thus creating a gold standard data set which can be used to train automatic recognizers for all sorts of constructions, including the "interesting ones", (i.e. construction-based parsers).

2.3 FN and MASC

FrameNet is serving as a subcontractor in the new project to create a multiply annotated subcorpus (MASC) of the American National Corpus. Because of the amount of text to be annotated, we will not be able to manually annotate more than a small portion, but will instead depend upon ASRL software, which we will retrain repeatedly as we add annotation during the course of the project.

A number of interoperability issues will have to

[1] An NSF SGER.

be solved for this task. For example, the FN annotation software has been set up such that, during the import process, all punctuation is separated from the words it occurs with by inserting a space. This very basic kind of tokenization simplifies a lot of manual annotation steps. But it also causes the pointers from the FN labels to the text to be offset by a few bytes here and there, depending on where the punctuation appears. The ANC staff have already agreed to post process the data to remove the offsets in order to integrate the FN data into the MASC. In the end, it may be necessary to rewrite the annotation software so that it will not depend on space-separated tokens. This would also have advantages for languages like Spanish that have enclitic pronouns. Also the different ASRL systems use different data formats (both for input and output), so some conversion scripts will need to be written to allow us to run them in parallel and compare their outputs.

2.4 Aligning WordNet and FrameNet

Many people have noted that WordNet has extensive lexical coverage, but minimal syntactic/valence information, while FrameNet has a rather limited lexicon, but quite detailed valence information about those lexical units. Staff at FrameNet and at WordNet are now working on a pilot study to align WordNet and FrameNet[2]. Since the two lexica were created for quite different purposes and have totally different data structures, the interoperability problems are obvious. This project will be discussed in a separate paper at this conference, so I will not go into it in detail here.

2.5 Development of image schema frames and their use in inference

A recent research direction has been an investigation of the frames needed to represent the image schemas described in current versions of Cognitive Linguistics. Properly representing image schemas is an essential step to being able to make the correct inferences from text. For example, in the Wikipedia sentence "…most indigenous peoples of the Americas descended from people who probably migrated from Siberia across the Bering Strait, anywhere between 9,000 and 50,000 years ago," image schemas provide much of the information needed to make the right inferences: In addition to the literal image schema of Siberia and the Americas with the Bering Strait between them, and a physical movement of people from one to the other, there are two metaphorical image schemas, one for the location of this event in time between 9,000 and 50,000 years ago, and the other viewing the change from one generation to the next as a movement downward, a *decent*. The words *descend, from, across*, and *between* all evoke image schemas which combine to create most of the meaning of the sentence, and allow us to infer many facts which are not explicit: that the ancestors of the peoples of the Americas were in Siberia before 50,000 years ago, that they completed this movement by 9,000 years ago, that Siberia, the Bering Strait, and the Americas are connected linearly in that sequence, etc.

We are planning to add roughly 100 frames in this area, which will cover many common prepositions and adjectives. We expect that doing so will require some realignment of existing frames, but will also greatly improve the FN semantic representation for much of everyday language.

3 FrameNets across Languages

SALSA project The SALSA project, based at Saarland University, and under the direction of Prof. Manfred Pinkal, has manually annotated the verbs in a German text, using their own very graphical annotation software (Erk et al., 2003), but applying the FN semantic frames and FEs so far as possible. The text is from the TIGER corpus, a parsed and manually corrected newswire corpus. Where they found no appropriate English frame, they created a verb-specific frame (something like PropBank) and simply called the FEs "FE1", "FE2", etc. They have just had their first data release.

The SALSA/TIGER XML format is quite different from any Berkeley FN XML format, as it is closely tied to a parse tree, which must be present for each sentence. In this respect, their work is like PropBank, since they are depending on the correctness of the parses in a TreeBank which has already been carefully manually validated (and which they themselves

[2]NSF IIS-0705155

helped validate). Katrin Erk and Sebstian Padó have also created a set of tools for Automatic Semantic Role Labeling collectively known as Shalmaneser (Erk and Padó, 2006) and released it to the world. It comes with pre-trained parameter sets for English and German; the English training was accomplished by first converting the FN annotation XML (LU XML files) to TIGER/SALSA XML.

Spanish FrameNet, based at Universidad Autónoma de Barcelona, under the direction of Prof. Carlos Subirats, is a lexicographic project which follows very closely the model of Berkeley FrameNet, using the same type of database and annotation software. They are even keeping up with version changes in the annotation software. The only part they have had to change was a small set of methods that attempts to guess the phrase type and grammatical function of phrases that have been annotated as FEs, since Spanish has different parts of speech and phrase types, and these changes have been brought back into the FN software distribution.

Spanish FN is extracting example sentences from their own corpus and formatting them in XML so that they can be imported by the same tools that import English sentences. They were also able to seed their tables of lexemes and word-forms from an existing Spanish lexicon, which greatly speeded up getting started on a lexicon. The Berkeley FrameNet lexicon was seeded in the same way from the CELEX lexicon, giving us roughly 40,000 lexemes whose word-forms we don't have to type in.

As with the SALSA project, the SFN team have found that some LUs in Spanish don't seem to fit into any of the English frames. Unlike SALSA, they are creating new frames and modifying existing frames as they go along, using the FN tools for this purpose. They are also using the FN report software to produce HTML and XML versions of their data, which will be released for the first time in the Spring of 2008. As this data is identical in XML format to the Berkeley FN data, it has been comparatively easy for our collaborator, Prof. Hiroaki Sato of Senshu University, Kawasaki, Japan to extend his FrameSQL website to produce comparisons of English and Spanish data, aligned according to a combination of frame identity across languages and translation equivalence between LUs.

Japanese FrameNet This project is the work of a group of scholars at Keio University and University of Tokyo, headed by Prof. Kyoko Ohara. They are using a modified version of the FN software to do annotation, but the changes seem to have been more extensive than for Spanish.

Chinese FrameNet This project is headed by Prof. Liu Kaiying of Shanxi University CS Dept and You Liping. Our information about this project is being based mainly on two conference papers, but it appears that they have built their own annotation software; but are using FrameNet frames to a great extent.

Portuguese FrameNet This is a new project that is just getting started in Minas Gerais, Brazil, as part of a collaboration agreement with ICSI. Prof. Subirats has visited the prospective team members and advised them of his experience in initializing the FN database for a new language.

All of the cross-linguistic FN projects raise the underlying question of how similar the frames of one language are to those of another; our initial impression is that a large portion of frames are indeed substantially the same across many languages. As there is a separate poster discussing this question, I will not go into it in detail here.

4 Toward a more collaborative FN

Finally, a word about some longer-range plans. It is becoming clear that progress on the FN lexicon will be slow if we continue to depend on expert manual annotation. We are investigating various ways of allowing a larger community of people to be involved in the day-to-day work of FrameNet. This is a delicate matter, as we want as wide participation as possible, but we want anything that is added to be

coherent with existing data and in accord with the annotation policy.

We want to make it possible for different people to participate in different ways, according to their interests and abilities. This will probably mean that the final word about what new frames are made and what LUs go in them will be coming from Berkeley, but that users of the data will be able to influence the direction of development, bring in texts for annotation, and actually do a lot of the annotation work. Exactly how this will be accomplished is under study; we are looking at a variety of models, including the Open Mind project and several data-collection on-line games, Wikipedia, the Stanford Encyclopedia of Philosophy, and Amazon.com's "Mechanical Turk", which would facilitate paying people (a small amount) for participating.

We also want to foster greater cooperation in the development of FrameNets in new languages, and to encourage specialists in particular domains to undertake work on the frames and LUs in those areas.

References

Katrin Erk and Sebastian Padó. 2006. Shalmaneser – a flexible toolbox for semantic role assignment. In *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Katrin Erk, Andrea Kowalski, and Sebastian Padó. 2003. The SALSA annotation tool. In *Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface*, Nancy, France.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language*, 64(1):501–538.

Charles J. Fillmore. 1968. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart & Winston, New York.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Paul Kay and Charles J. Fillmore. 1999. Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction. *Language*, 75:1–33.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. July 2004. The Sketch Engine. In *Proceedings of EURALEX 2004*, Lorient, France.

Enhancing Czech Valency Lexicon with Semantic Information from FrameNet: The Case of Communication Verbs

Václava Benešová, Markéta Lopatková and Klára Hrstková
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Malostranské nám. 25, Prague
{benesova,lopatkova,hrstkova}@ufal.mff.cuni.cz

Abstract

In this paper, we report on our attempt at assigning semantic information from FrameNet to lexical units in VALLEX, a valency lexicon of Czech verbs. We focus on the class of communication verbs. We experiment with assigning FrameNet semantic frames to lexical units for communication verbs. The second task consists in linking their valency complementations with FrameNet frame elements. The exact pairwise inter-annotator agreement reaches almost 69% on the semantic frames and 84.6% on the frame elements. We propose enhancing VALLEX with missing semantic information from FrameNet based on exploitation of the semantic relation 'Inheritance'.

1 Introduction

Syntactic and semantic behavior of verbs, which are considered to be the syntactic centers of sentences, is the key information for any rule-based tasks of NLP. Apparently, various theoretical approaches to valency are reflected in the annotation schemes of particular lexical resources (for a short survey and characteristics of the most important projects, see e.g. (Žabokrtský, 2005)). For linguists, their dissimilarity represents a challenging task of their comparison and mutual enriching.

In this article, we describe an experiment with enhancing VALLEX, a valency lexicon of Czech verbs (Žabokrtský et al., 2007), with FrameNet semantic information. In the experiment, we attempted at linking FrameNet (Baker et al., 1998) semantic features – semantic frames and frame elements – to the VALLEX data. First, lexical units in VALLEX were assigned semantic frames. Then their valency complementations were linked with frame elements.

VALLEX, using the Functional Generative Description as its theoretical background, see (Sgall et al., 1986), takes into account mainly the syntactic criteria for identifying arguments. However, the semantic information is crucial for handling such NLP tasks as generation, information retrieval, or question answering. Therefore, we have decided to obtain this missing information from the Berkeley FrameNet project. In spite of its being still in progress, it represents an elaborated semantically oriented lexical resource describing valency for English.

We exploit the tentative classification of verbs in VALLEX, sorting verbs into rough and heterogeneous 'supergroups'. We pursue three goals: First, we attempt at further structuring a particular group of verbs into more homogenous, subtler classes. Second, we expect to observe a hierarchy of semantic relations between Czech verbs analogous to the hierarchy of English verbs[1] as a result of assigning semantic frames to a semantically coherent group of Czech verbs. Last, we aim at choosing coarse-grained semantic information from FrameNet for enhancing VALLEX.

The paper is structured as follows: in Section 2, we present basic properties of VALLEX and FrameNet, Section 3 describes the method of our ex-

[1] Displayed by Frame Grapher Tool, see http://framenet.icsi.berkeley.edu/FrameGrapher/

periment with assigning semantic information from FrameNet to lexical units from VALLEX. Section 4 provides the analysis of the material obtained from the experiment. Lastly, the outline of further exploiting FrameNet semantic information for VALLEX follows in Section 5.

2 A Brief Characterization of Two Annotation Schemes: VALLEX and FrameNet

In this section, we briefly describe two approaches to the description of valency: VALLEX, which takes into account mainly syntactic criteria for identifying arguments, and semantically oriented FrameNet.

2.1 Valency Lexicon of Czech Verbs: VALLEX

The Valency Lexicon of Czech Verbs, VALLEX, version 2.0[2], provides information on the valency structure of verbs in their particular senses, lexical units (LUs).

LUs are considered, identically with (Cruse, 1986), as "form-meaning complexes with (relatively) stable and discrete semantic properties"; roughly speaking, 'the given word in the given sense'. Distinguishing senses is based on both syntactic and semantic properties (i.e., considerable shifting in meaning).

Each LU covers both the perfective and the imperfective Czech verbs (if they exist) that create a single lexeme. The following Table 1 shows the basic statistics about VALLEX.

	VALLEX 2.0
Number of lexeme entries	2731
Number of verbs	4250
Number of LUs	6462
Number of LUs with a class	2903
Number of classes	22
Lexical units with a class (%)	44.9%

Table 1: Basic statistics about VALLEX 2.0

Valency Frames. In VALLEX, the key information on the valency structure of a given LU is encoded in the form of *valency frames*. Valency frames are formed as a sequence of slots; each slot stands for one valency complementation and consists of its type ('Actor', 'Addressee', etc.), morphemic realization and its obligatoriness (obligatory or optional), see below. An example of the lexeme for the verbs *doplňovat*^{impf}, *doplnit*^{pf} 'to add' as captured in VALLEX can be seen in Figure 1.

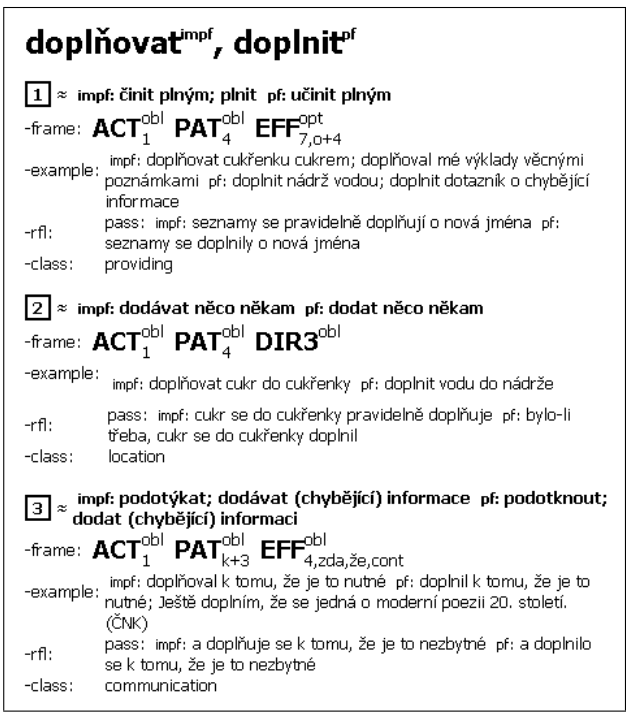


Figure 1: The lexeme for the verbs *doplňovat*^{impf}, *doplnit*^{pf} 'to add' in VALLEX, consisting of three LUs.

Valency Complementations. In VALLEX, based on the *Functional Generative Description* (FGD, see esp. (Sgall et al., 1986), (Panevová, 1974)), valency complementations (VCs) are divided into arguments (inner participants) and free modifications (adjuncts). They both can be obligatory or optional.

(Verbal) arguments are distinguished rather on the basis of the syntactic behavior of verbs. Two criteria are applied (introduced in (Panevová, 1974)):

- each argument can modify only a more or less closed class of verbs (that can be listed),
- each argument can modify a particular verb only once (except for the case of coordination).

[2] http://ufal.mff.cuni.cz/vallex /2.0/

Moreover, possible morphemic realization(s) of these arguments is/are typically determined by the governing verb.

Five types of arguments have been determined – 'Actor' (label ACT), 'Patient' (PAT), 'Addressee' (ADDR), 'Origin' (ORIG), and 'Effect' (EFF).

According to FGD, see (Panevová, 1974), if a verb has only one argument, then this argument is 'Actor' and if a verb has two arguments, then these are 'Actor' and 'Patient'. PropBank (Palmer, 2005) has the similar approach at least to non-ergative verbs.[3] When determining the other three arguments, semantic criteria are also taken into account. As a consequence, the types of arguments do not always reflect the exact semantic relation between a verb and its arguments.

In contrast to the arguments, the *free modifications* are semantically distinctive, being identified on the basis of their syntactico-semantic functions.

Verb Classes in VALLEX. At the present time, a selected part of LUs in VALLEX are assigned semantic classes like 'motion', 'communication', 'perception', etc. These classes were built in a 'bottom-up' way mainly on the basis of the syntactic properties of LUs, with regard to their semantics. These rough heterogeneous 'supergroups' – although not based on a properly defined ontology – can represent an efficient starting point for building an applicable semantic classification of verbs.

Communication verbs. The reported project focuses on the 'communication verbs' (the so-called 'verba dicendi'). These verbs can generally be specified as verbs rendering situation when 'a speaker conveys information to a recipient'. They are characteristic by the entity of 'information' which can be expressed as a dependent content clause.

This class of verbs was convenient for our experiment for two reasons: first, they have specific syntactic behavior, and second, they represent a good-sized (large enough) class. Moreover, we assume that the results will be applicable also to other groups of verbs with a sentential complement.

[3]However, the status of ergative verbs in the Slavic languages are not clear.

2.2 FrameNet

The FrameNet lexical database[4] is an on-line lexical resource for English, see (Baker et al., 1998). Its aim is "to document the range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses, through a computer-assisted annotation of example sentences", see (Ruppenhofer et al., 1986).

As to the quantitative characteristics, FrameNet contains more than 10 thousand lexical units (a pair consisting of a word and its meaning)[5] in more than 825 semantic frames, exemplified by around 135 thousand annotated sentences. At present, the project focuses primarily on verbs, nouns and adjectives.

Semantic Frames. The descriptive framework of FrameNet is based on *frame semantics*. Each LU evokes a particular semantic frame underlying its meaning. Each *semantic frame* (SF) can be understood as a "conceptual structure describing a particular type of situation, object, or event", see (Ruppenhofer et al., 1986). Each SF contains the so-called frame elements, i.e., semantic participants which are seen as components of such situations.

For example, the SF 'Statement' is defined as follows: "This frame contains verbs and nouns that communicate the act of a Speaker to address a Message to some Addressee using language ...", see FrameNet webpage.

Frame Elements. Semantic frames consist of *frame elements* (FEs), semantic arguments of a predicating word evoking this frame.

Whereas, for instance, *Case Grammar*, see (Fillmore, 1968) assumes a fixed, relevant-across-the-board collection of underlying 'cases', FEs representing semantic information are understood in terms of roles in specific frames and not as a restricted set of universal semantic roles. It implies that the inventory of FEs is specific for each SF.

Three types of FEs are distinguished, *core FEs* (conceptually necessary FEs whose combination is characteristic for a particular SF), *peripheral FEs* (not unique for a given SF, they can occur in any

[4]http://framenet.icsi.berkeley.edu/

[5]We use the same abbreviation LU both for VALLEX and FrameNet because the same concepts are concerned in principle.

SFs) and *extra-thematic FEs* (that set a given event on the background of another event or state of the same type). E.g., the SF 'Statement' consists of core FEs 'Speaker', 'Topic', 'Message' and 'Medium' and peripheral FEs 'Addressee', 'Depictive', 'Degree', 'Epistemic_stance', 'Group', 'Internal_cause', 'Manner', 'Means', 'Occasion', 'Particular_iteration', 'Place' and 'Time'.

The following sentence illustrates the SF 'Statement' and its FEs:

> *President Kennedy*.Speaker *said to an astronaut*.Addressee: *("Man is still the most extraordinary computer of all.")*.Message.

Hierarchy of Semantic Relations between SFs. FrameNet builds a wide network of hierarchical relations between SFs and their FEs. The most important relations are the following, see (Ruppenhofer et al., 1986):

- 'Inheritance' – everything which is true about the semantics of the parent frame holds for the semantics of its child frame(s). Each FE from the parent frame (except for extra-thematic FEs) is related to a relevant FE in the child frame.

- 'Using' – the parent frame constitutes the background for its child frames. Not all FEs from the parent frame must be bound to the FEs from the child frame.

- 'Subframe' – the child frame instantiates a part of a complex event represented by the parent frame.

For the purpose of enhancing VALLEX with semantic information, we exploit the transitive relation 'Inheritance', as will be discussed in Section 5.

3 Assigning Semantic Information from FrameNet to Valency Frames in VALLEX

In this section, we report on assigning the semantic information from FrameNet to the VALLEX communication verbs. In the first step, we translated each LU from Czech to English.[6] The total number of translated Czech LUs was 341 (without idiomatic LUs). These LUs correspond to 531 Czech verbs, counting perfective and imperfective verbs separately.

3.1 Assigning Semantic Frames and Frame Elements

Two human annotators (referred A_1 and A_2 in the sequel) searched each translated English LU in FrameNet and indicated an appropriate SF (labeled as 'Unambiguous Annotations'). The annotators were allowed to assign more than one SF to a particular LU ('Ambiguous annotations') – if the English equivalents belonged to more than one SF.

In two situations, the annotators could conclude that the given English LU is missing in FrameNet: (i) the corresponding lemma was missing in FrameNet at all, or (ii) the found English LU did not correspond to the meaning of the given Czech LU. The following Table 2 shows the basic statistics concerning assigning SFs.

	A_1	A_2
Cz LUs	341	341
Eng equivalents	653	653
Annotations of SFs	610	556
Unambiguous annotations of SFs	143	165
Ambiguous annotations of SFs	467	391
Marked as missing Eng SFs	11	19
Marked as missing Eng LUs	33	35

Table 2: Annotated data size and statistics of two annotations of SFs.

After having indicated the appropriate SF(s), the annotators had to assign the corresponding FE(s) of this/these SF(s) to each valency complementation (VC) of the given Czech LU. Similarly as in the case of assigning SFs, a valid answer indicated appropriate FE(s) ('Unambiguous/Ambiguous annotations'). In the cases when no suitable FE was found, the annotators used a special flag. Table 3 gives the numbers of VCs and FEs used in the experiment.

[6] The on-line dictionary at http://www.lingea.cz/ was used for manual translation.

	A_1	A_2
Annotations of VCs from VALLEX	1088	1088
Annotations of FEs	1322	1314
Unambiguous annotations of FEs	869	879
Ambiguous annotations of FEs	453	435
Marked as missing Eng FEs	47	34

Table 3: Annotated data size and statistics about two annotations of FEs.

3.2 Results: Inter-annotator Agreement

Table 4 summarizes the inter-annotator agreement (IAA) and Cohen's κ statistics, see (Carletta, 1996). The exact match of answers relating to SFs reaches 68.8%. The κ statistics compensates IAA for agreement by chance. The level relating to SFs that we achieved (0.47) represents a very moderate agreement, see (Krippendorff, 1980). However, the intersection match (if both annotators chose the same SFs regardless of other variants in the case of ambiguous annotations) gives a more satisfactory result (88.2%, κ = 0.79). IAA relating to FEs is measured only in cases of an exact match of SFs (401 cases). IAA concerning FEs (84.6%, κ = 0.83) is much better in comparison with SFs. The intersection match concerning FEs represents a significant result (93.3%, κ = 0.92).

	IAA [%]	κ
Exact match of SFs	68,8%	0.47
Intersection match of SFs	88,2%	0.79
Exact match of FEs	84,6%	0.83
Intersection match of FEs	93,3%	0.92

Table 4: Inter-annotator agreement and κ statistics.

4 Analysis of Obtained Material

In this section, we describe the analysis of the material obtained from our experiment mainly from the linguistic point of view. Special attention is paid to ambiguous assignment of SFs to Czech LUs and FEs to valency complementations.

4.1 Analysis of Assigned Semantic Frames

The annotators assigned 100 SFs from FrameNet to 341 communication verbs from VALLEX. The following SFs belong to the most often assigned:

- 'Statement' (141 cases in 2 annotations), e.g., *dodat*^{pf} 'to add', *oznámit*^{pf} 'to announce', *poznamenat*^{pf} 'to remark', *sdělit*^{pf} 'to tell', ... ,
- 'Request' (76), e.g., *nakázat*^{pf} 'to order', *naléhat*^{impf} 'to urge', *žádat*^{impf} 'to plead', ... ,
- 'Telling' (59), e.g., *povědět*^{pf} 'to tell', *říci*^{pf} 'to say', ... ,
- 'Communication_manner' (35), e.g., *křičet*^{impf} 'to shout', *šeptat*^{impf} 'to whisper', *zamumlat*^{impf} 'to gabble', ... ,
- 'Reporting' (34), e.g., *nahlásit*^{pf} 'to inform', *udat*^{pf} 'to report', ... ,
- 'Attempt_suasion' (31), e.g., *povzbudit*^{pf} 'to encourage', *vybízet*^{impf} 'to urge', ...

Ambiguous Assignment of Semantic Frames. From the linguistic point of view, the cases when the annotators assigned two or more SFs to one Czech LU are the most interesting, as in the following scheme:

$$\text{Cz LU} \begin{cases} \nearrow \text{Eng LU}^{I} \longrightarrow \text{SF}^{I} \\ \searrow \text{Eng LU}^{II} \longrightarrow \text{SF}^{II} \end{cases}$$

The SFs which were systematically assigned ambiguously to Czech LUs refer to regular differences which FrameNet and VALLEX make in word sense disambiguation.

Describing valency frames, VALLEX leaves aside in/animateness of the entities occupying one valency position, in contrast to semantically based FrameNet. As a result, some Czech verbs represented by one LU in VALLEX belong to two (or even more) LUs in FrameNet.

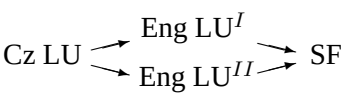
For instance, the following instances of the verb *dokázat*^{pf}, *dokazovat*^{impf} 'to prove' are described by one valency frame in VALLEX.

> *Peter has proved that the given solution was out-of-date.*
> *The facts have proved that the given solution was out-of-date.*

In FrameNet, the first instance corresponds to English LU^{I}, 'to prove' from SF^{I} 'Reasoning', whereas the second instance belongs to English LU^{II}, 'to prove' from SF^{II}'Evidence'.

This fact concerns a number of other n-tuples of SFs, e.g., the SF 'Grant_permission' – 'Permitting', or 'Judgment_communication' – 'Judgment' – 'Notification_of_charges', etc.

In several cases, Czech verbs had more different translations belonging to the same SF, as in the following scheme.



In contrast to the above mentioned ambiguous assignment of SFs, we do not consider these cases interesting.

4.2 Analysis of Assigned Frame Elements

The annotators assigned 116 types of FEs to valency complementations. The most often assigned FEs are 'Speaker' (545 times in 2 annotations), 'Addressee' (485), 'Message' (393), 'Medium' (358), 'Topic' (330), 'Communicator' (92), 'Content' (81), etc.

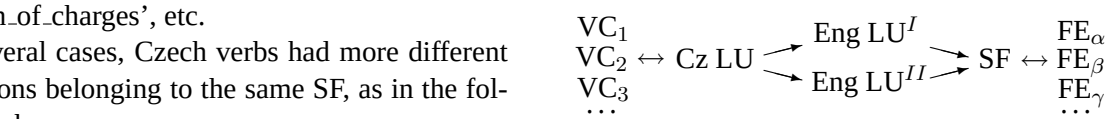
Logically, the most often assigned FEs come from the most frequently assigned SFs. Moreover, some FEs are parts of more than one SF: 'Speaker' belongs to the SFs 'Statement', 'Telling', 'Request', 'Communication_manner', etc.

After having assigned FEs to the arguments of communication verbs, we can summarize which semantic properties are relevant for entities occupying their valency positions. The most significant are the following:

- 'Actor' matches esp. 'Speaker', 'Medium', 'Communicator', 'Voice', 'Author',
- 'Patient' matches esp. 'Topic', 'Message', 'Content',
- 'Addressee' matches esp. 'Authorities', 'Evaluee', 'Grantee'.

We can observe that both the animate (e.g., 'Speaker') and the inanimate (e.g., 'Voice') entities can occupy the positions of 'Actor' by communication verbs. Predominantly animate entities play the role of 'Addressee'. 'Patient' and 'Effect' are characterized by abstract semantic properties.

Ambiguous Assignment of Frame Elements. We observe two types of ambiguous assignment of FEs. The first type occurs if a particular Czech LU has only one English equivalent or it has more translations but all of them belong to one SF (see Section 4.1), as in the following scheme.



Then there may be more FEs from this SF that correspond to a particular valency complementation. We can illustrate this case with the verb *děkovat*^{impf} translated as 'to thank', which belongs to the (only one) SF 'Judgment_direct_address' but has an ambiguous assignment of FEs.

VCs	corresponding FEs
Actor	Communicator, Medium
Addressee	Addressee
Patient	Reason, Topic

This case of the ambiguous assignment of FEs often results from the different approach to in/animateness which FrameNet and VALLEX have. As VALLEX does not take into account in/animateness of the first and second arguments, 'Actor' and 'Patient' are often assigned ambiguously in contrast to 'Addressee' and 'Effect'.

The ambiguous assignment of FEs to 'Patient' often follows from the fact that one abstract entity can express both 'theme' and 'what is said about the theme' by Czech communication verbs, as in the following sentence:

> *The news talked (about the horrible earthquake that struck Turkey on Friday morning).*Topic, Message).

Moreover, both 'Topic' and 'Message' can be realized separately in one Czech sentence, as in the following example, see also (Daneš et al., 1987), (Panevová, 1974):

> *Cizinci si stěžují starostovi na obchodníky.*Topic, *(že užívají dvojí ceny).*Message.
> 'foreigners – refl – complain – city_mayor – about – sellers – that – use – double – prices'
> Eng. *The foreigners complain to the city mayor that the sellers use double prices.*

The second type of the ambiguous assignment of FEs is closely related to the ambiguous assignment

of SFs (see Section 4.1). If one Czech LU was assigned more than one SF, then the valency complementations of such Czech LU were assigned FEs from all these SFs. Therefore, the ambiguous assignment of FEs automatically arises from the ambiguous assignment of SFs.

For instance, one Czech LU *dokázat*^{pf}, *dokazovat*^{impf} translated as 'to prove' corresponds to two SF^{I} 'Reasoning' and SF^{II} 'Evidence' in FrameNet (see Section 4.1) – as a result, the valency complementations of this LU are assigned FEs from both SFs. In these cases, all valency complementations can be assigned more than one FE:

VCs	FEs from SF^{I}	FEs from SF^{II}
Actor	Arguer	Support
Addressee	Addressee	Cognizer
Patient	Content	Proposition

5 Exploiting Semantic Information from FrameNet for VALLEX

In this section, we outline a further exploitation of the semantic information from FrameNet for VALLEX. We propose enhancing VALLEX with coarse-grained semantic information based on the semantic relation 'Inheritance'.

5.1 Structuring Semantic Classes in VALLEX.

FrameNet distinguishes several types of semantic relations, on the basis of which the semantic information is provided on different levels of granularity (see Section 2.2).

We consider the semantic relation 'Inheritance' as the most important. The child frame, although more specific than its parent frame, inherits all semantic properties from it. This concerns FEs, their relations to each other and frame-to-frame relations.

We exploit the semantic information from the top levels of the 'Inheritance' relations. This method allows classifying 'supergroups' of communication verbs in VALLEX into well-defined, coarse-grained classes from FrameNet as 'Communication', 'Prohibiting', 'Judgment_communication', etc., see also Figure 2.

We have obtained 59 top SFs for Czech communication verbs. (However, more than a half of the total number of these SFs have not been integrated into the network of the relation 'Inheritance' yet. We suppose continuously complementing the top levels of SFs in the future. Therefore, the final number of coarse-grained semantic classes is assumed to be significantly lower.)

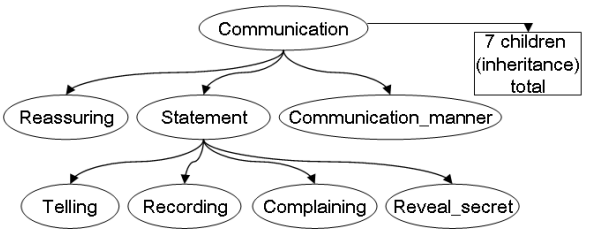


Figure 2: The relation of 'Inheritance' in FrameNet for the SF 'Communication'. (The SF 'Communication' represents the top SF in the relation 'Inheritance', e.g., for the SFs 'Reassuring', 'Communication_manner' and 'Statement', and transitively also for their children: 'Telling', 'Complaining', etc.)

5.2 Exploitation of Top Frame Elements as Semantic Roles.

Each FE from a child frame represents a subtype of the corresponding FE in its parent SF. Thus it allows us to assign FEs from the top SF in the relation 'Inheritance' to valency complementations of Czech communication verbs which were assigned SFs from the lower levels in this relation. Figure 3 shows the relevant relations between FEs of the SFs 'Communication', 'Statement' and 'Telling'.

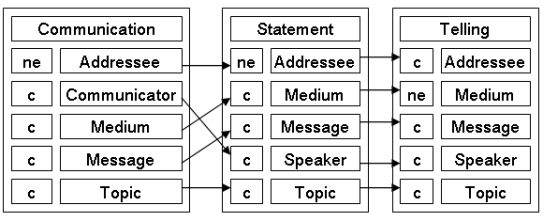


Figure 3: The FEs-to-FEs relations between the SFs 'Communication', 'Statement' and 'Telling' relevant for the valency complementations of Czech communication verbs assigned with these SFs.

Exploiting FEs as semantic roles from the top levels of 'Inheritance', we enrich the lexicon with more general, coarse-grained but extensive FEs. They provide VALLEX lexicon with sufficient informa-

tion on the selectional preferences of the individual arguments. The following example shows assignment of FEs to valency complementations of the verb *informovat*^{biasp} 'to inform' from the SF 'Telling', a subtype of the SF 'Communication':

> *Učitel.*ACT-Communicator *informoval rodiče.*ADDR-Addressee, *(že jejich syn má špatné známky).*PAT-Message.
> Eng. *The teacher has informed the parents that their son has bad marks.*
>
> *Noviny.*ACT-Medium *informovaly čtenáře.*ADDR-Addressee, *(že ve věznici panují otřesné podmínky).*PAT-Message.
> Eng. *The newspapers have informed readers that outrageous conditions reign in the prison.*

Therefore, the enhanced valency frame for the corresponding LU for the verb *informovat*^{biasp} 'to inform' has the following form:

VCs	corresponding FEs
ACT	Communicator | Medium
ADDR	Addressee
PAT	Topic
EFF	Message

6 Conclusions and Future Work

We have presented an experiment in which VALLEX data were assigned semantic frames and frame elements from FrameNet. We attained a satisfactory inter-annotator agreement, especially concerning the FEs. We have proposed a method of enhancing VALLEX with the semantic information from FrameNet based on the relation 'Inheritance'. Focusing on communication verbs, we obtained applicable top level hierarchies.

For future work, we plan to assign the semantic information from FrameNet to other verbs with a sentential complement, namely to the verbs included in the VALLEX classes 'mental action', 'psych verb' and 'perception'. The significant inter-annotator agreement on assigning FEs promises good results for (semi)automatic assigning FEs as semantic roles.

Acknowledgement. The research reported in this paper is carried under the project of the Ministry of Education, Youth and Sports of Czech Republic No. MSM0021620838, under the grants LC536 and GA UK 7982/2007.

References

Collin F. Baker, Charles J. Fillmore, John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Jean Carletta. 1996. Assessing agreement on classification task: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

D. Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

František Daneš and Zdeněk Hlavsa et al. 1987. *Větné vzorce v češtině*. (2nd ed.) Academia, Praha.

Charles J. Fillmore. 1968. *The Case for Case*. In Bach, E., Harms R.T., eds. *Universals in Linguistic Theory*. Holt, Rinehart and Winston, 1–88.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Beverly Hills CA: Sage.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Jarmila Panevová. 1974. On Verbal Frames in Functional Generative Description. *The Prague Bulletin of Mathematical Linguistics*, 22:3–40.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher Johnson and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. http://framenet.icsi.berkeley.edu/book/book.html/.

Petr Sgall, Eva Hajičová and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel Publishing Company, Dordrecht.

Zdeněk Žabokrtský. 2005. *Valency Lexicon of Czech Verbs*, Ph.D. thesis. Faculty of Mathematics and Physics, Charles University, Prague.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. Valency Information in VALLEX 2.0. Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics*, 87:41–60.

Fully Embedded Type Systems for the Semantic Annotation Layer

Ekaterina Buyko & Udo Hahn
Jena University Language & Information Engineering (JULIE) Lab
Fürstengraben 30, 07743 Jena, Germany
{buyko|hahn}@coling-uni-jena.de

Abstract

We extend an integrated linguistic annotation type system which already covers the formal document layer, lexical and morpho-syntactic linguistic specifications by a semantic layer for annotation types. The extension is exemplified using the fairly limited set of semantic entities from ACE-style newspaper language domain annotation efforts. We also suggest a solution for those domains where the number of semantic entities is usually several orders of magnitude larger and, hence, much more diverse. As an example, we deal with the life sciences domain where we define a clean interface between generic semantic annotation types and the highly elaborated conceptual structure of bio domain ontologies. Furthermore, we demonstrate by means of comprehensive use cases the benefits of merging various linguistic annotation layers under the umbrella of an integrated type system.

1 Introduction

Human language technology development and natural language engineering activities, in the past years, have led to the creation of considerable amounts of linguistic meta data at virtually all linguistic levels – ranging from lexical to morpho-syntactic and semantic specifications. This work has usually focused on linguistic annotations only from a single-level perspective, e.g., POS tags, phrase structure descriptions, or semantic specifications. Recently, however, the NLP community shifted its attention to combine and merge different kinds of annotations for single or even for multiple annotation layers (cf., e.g., the activities reported by the *Annotation Compatibility Group* (Meyers, 2006)).

From a system engineering perspective, a common specification framework that would connect different levels of linguistic analysis is recognized as crucial for successfully combining and merging linguistic annotations (cf., e.g., Bird and Liberman (2001), Ide et al. (2003), Ferrucci and Lally (2004)).

In this paper, on the one hand, we want to provide specifications for semantic annotations where various proposals can be integrated within a generic framework in a coherent way, in addition to already integrated linguistic annotation layers (e.g, morphology or syntax). On the other hand, given the formal potential of semantic type hierarchies (e.g., feature inheritance), we want to sort out commonalities they share so that the increasing levels of abstractions they implicitly contain can be made explicit. In particular, we propose a clean demarcation line in terms of an interface definition for large-coverage lexicons or ontologies which hook up to generic semantic annotation units. This strategy will allow us to avoid the duplication of information held in different language resources (lexicons, ontologies, etc.) at the more abstract level of semantic meta data.

2 Related work

The design of annotation schemata for language resources and their standardization is often limited to single layers of linguistic analysis, e.g., syntactic or semantic analysis. Syntactic annotation schemata,

such as the one from the Penn Treebank (Marcus et al., 1993), or semantic annotations, such as those underlying ACE (Automatic Content Extraction Program) (Doddington et al., 2004) or TIMEML (Pustejovsky et al., 2003a), are increasingly considered as a *de facto* standard.

Recently, however, the NLP community has started to combine different kinds of linguistic annotations. Major bio-medical corpora, such as GENIA (Ohta et al., 2002) or PennBioIE,[1] incorporate several layers of linguistic information in terms of morpho-syntactic, syntactic and semantic annotations within one corpus.

Assembling various annotation levels, however, did not result in an annotation scheme for a complete NLP pipeline as needed, e.g., for information extraction or text mining tasks. This lack was mainly due to missing standards for specifying comprehensive NLP software architectures. Proposals of annotation schemata for a complete NLP pipeline usually shared their explicit linkage to a specific NLP tool suite or NLP system and thus suffer from the provision of a generic annotation framework that can be re-used in other developmental activities (Buitelaar et al. (2003), Schäfer (2006)).

Recently started international standardization frameworks such as LAF under the auspices of the ISO TC37 SC WG-1-1 (Ide et al., 2003), and the emergence of generic NLP frameworks such as UIMA (Unstructured Information Management Architecture) (Ferrucci and Lally, 2004)), reflect the community's intention to create such integrated annotation schemata. Several annotation schemata (*type systems* in UIMA jargon) were developed within the UIMA framework (Hahn et al. (2007a), Piao et al. (2007)). These type systems indeed do cover the results of linguistic pre-processing, as well as document meta and structure analysis, but they still lack more elaborate specifications at the semantic level of annotation (e.g., relations and events). Furthermore, it is not at all clear how these type systems could be extended when already existing semantic annotation schemata or terminologies come into the play.

Welty and Murdock (2006) introduced in their work HUTT (Hierarchical Unified Type Taxonomy), an UIMA type system integrating a variety of established information extraction taxonomies (e.g, ACE (Doddington et al., 2004), TIMEBANK (Pustejovsky et al., 2003b), etc.).

As an alternative, we use existing generic (*core*) type systems as a backbone for further definitions and extensions aiming, e.g., at information extraction tasks. We integrate established semantic category systems and ontologies from the newswire and biomedical domains, respectively. Furthermore, we show by means of comprehensive use cases the extended type systems in action and demonstrate the benefits of combining various linguistic annotations under the umbrella of a generic type system.

3 Graph-based Annotation Models

The uniform representation of annotated data is recognized as crucial for the interoperability between various linguistic (e.g., morpho-syntactic, syntactic, semantic) annotation layers. In particular, graph-based annotation models were shown to be especially appropriate for linguistic annotation. They provide the scaffold for the design of annotation schemata. The implementation of annotation schemata is then realized within annotation frameworks implementing the model.

We first introduce the graph-based annotation model in a more detailed way and then turn to some annotation schemata based on this model.

3.1 Annotation Graphs

There are two basic concepts in the graph-based annotation model, *viz.* nodes and edges. *Nodes* represent feature structures providing the annotation content. Nodes are linked by *edges* either to the original subject of analysis or to other feature structures. According to these specifications, an annotation is a graph which represents a feature structure.

Ide and Suderman (2007) formalize the graph-based model as a part of the LAF, a general framework for representing annotations. A graph of annotations *G* is defined as a set of vertices *V(G)* and a set of edges *E(G)*. Vertices and edges can be labelled with features. A feature is characterized as a quadruple (*G, VE, K, V*), where *G* is a graph, *VE* is a vertex or edge in *G*, *K* is a feature name and *V* a feature value.

[1] `http://bioie.ldc.upenn.edu`

Frameworks implementing the graph-based annotation model are, e.g., LAF and UIMA. LAF provides the GRAF (Graph-based Format) (Ide and Suderman, 2007), an XML serialization format of the model, while UIMA comes with CAS (Common Analysis Structure), an object-oriented implementation of the model (Götz and Suhre, 2004). As UIMA provides a platform for the development and deployment of large-scale NLP applications, it is attracting more and more attention in the NLP community. In the following, we will focus on CAS, the data model of the UIMA framework.

3.2 UIMA Common Analysis Structure

Common Analysis Structure (CAS) is a part of the system that controls the data flow in the UIMA architecture. CASes contain the original subject of analysis and annotations in the format of CAS objects. An annotation associates one CAS object with a region in the subject of analysis (e.g., the start and the end positions in the document).

Each CAS object (so-called *type*) consists of *feature structures*. Features specify slots within types and can either have primitive values such as integers or strings, or reference other instances of types in a CAS. Types can be extended via the inheritance mechanism. All types in a CAS derive from the basic type UIMA.TCAS.ANNOTATION. and thus inherit the basic annotation features, *viz. begin* and *end* (referencing spans of annotations in the subject of analysis). Accordingly, CASes provide the platform for the design of annotation schemata (*annotation type systems* in the UIMA jargon).

3.3 Annotation Type Systems

The definition of proper annotation schemata (the definition of labels and their interrelations) is usually carried out within an annotation framework implementing the annotation model. Recently some annotation schemata (type systems) were developed within the UIMA Framework.

The generic type system of Hahn et al. (2007a) aims to provide a core domain-independent type system for linguistic annotation which is extensible to domain-specific type systems. We present here the basic concepts of this type system.

Multi-layered type system – The type system consists of five annotation layers which cover the annotation of analysis results from the main parts of an NLP pipeline.

- the *Document Meta* layer represents the bibliographical information of the document and the information about the document content;
- the *Document Structure & Style* layer represents the formal structure of the document, e.g., its division in text body and title, paragraphs and sections, etc.;
- the *Morpho-Syntax* layer contains types necessary for the annotation of text segmentation (e.g., tokenization), and stores the results of the morpho-syntactic analysis such as lemmatization, POS tagging, etc.;
- the *Syntax* layer consists of types for the representation of parsing results (shallow and full parsing, dependency- or constituency-based parsing);
- the *Semantics* layer currently provides only basic types for the annotation of named entities (with pending extensions to relations and events).

***Core* types** – Each layer contains *core* annotation types that are defined as domain-independent, e.g., `POSTag` at the *Morpho-Syntax* layer, `Constituent` at the *Syntax* layer, or `Entity` at the *Semantics* layer.

***Specialized* types** – The core types can then be extended by more specialized types. `POSTag`, e.g., might be extended by types that represent specific tag sets, e.g., `PennPOSTag` (Marcus et al., 1993) or `GeniaPOSTag` (Ohta et al., 2002). For the semantic types we find potential extensions to the biomedical domain in terms of types such as `Gene` or `Organism`.

Implementation in CAS – All feature structures are first defined framework-independently in UML (Unified Modelling Language).[2] The type system is implemented as an UIMA CAS type system.

In our work, we employ this type system as a backbone for further extensions. Our extensions affect the *Semantics* layer only. First, we define additional semantic core types needed for information

[2] `http://www.uml.org/`

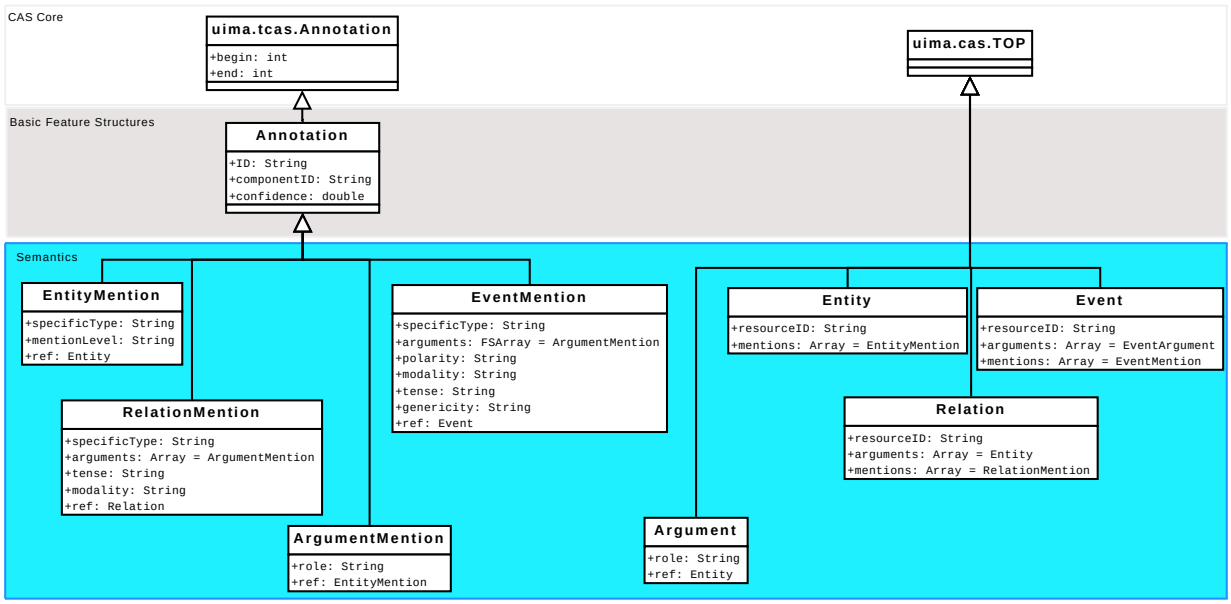


Figure 1: *Semantics* Layer in a Multi-layered UIMA Annotation Scheme in UML Representation

extraction applications, i.e., types representing relations and events. We then extend the core semantic types to the domain-specific types, in particular for the newswire and biomedical domain.

4 Core *Semantics* Layer — Fundamental Types

For information extraction tasks we extend the original *Semantics* layer of Hahn et al. (2007a) by introducing new core types, *viz.* `Relation`, `Event`, and `EventArgument` (see Figure 1). These new type definitions are related to already established semantic annotation schemata (cf., e.g., Doddington et al. (2004)).

We distinguish at the *Semantics* layer between types which refer to instances or states in the real world (e.g., `Entity`) and types which refer to linguistic data such as text spans that contain mentions of these objects (e.g., `EntityMention`). The CAS objects of the type `EntityMention`, `RelationMention`, `EventMention` which are coreferent are aggregated by the CAS object of the type `Entity`, `Relation` or `Event`, respectively. The *non-mention* types extend the `uima.cas.TOP` type and thus do not dispose of the features *begin* and *end* referring to the text span in contrast to the *mention* types which extend the `Annotation` type. The *non-mention* types enhance the `uima.cas.TOP` type with two additional features, *viz. mentions* and *resourceID*. The feature *mentions* refers to all mentions of the entity or relation in the text collection, while the feature *resourceID* refers to any resource entry which is unique for this entity (e.g., a pointer to an ontology or a gazetteer).

The `Mention` types enhance the `Annotation` type with two default features, *specificType* which specifies the type of the mention and usually refers to a terminology, and the feature *ref* referring to the instance of the entity, relation or event, accordingly. This default feature structure is supplied by the additional features specific for entity, relation and event, accordingly. In the following we describe the `Mention` types in a more detailed way.

`EntityMention` refers to the mentions of named entities in the text and provides an additional feature, *mentionLevel*, which specifies the mention level, e.g., *pronominal* or *nominal*.

`RelationMention` stands for a binary semantic relation between entities. The relation is considered as an ordered pair of entities. The type `ArgumentMention` represents the argument of the relation and refers through the feature *ref* to the particular `EntityMention`. The feature *role* assigns the role to the entity in the relation under

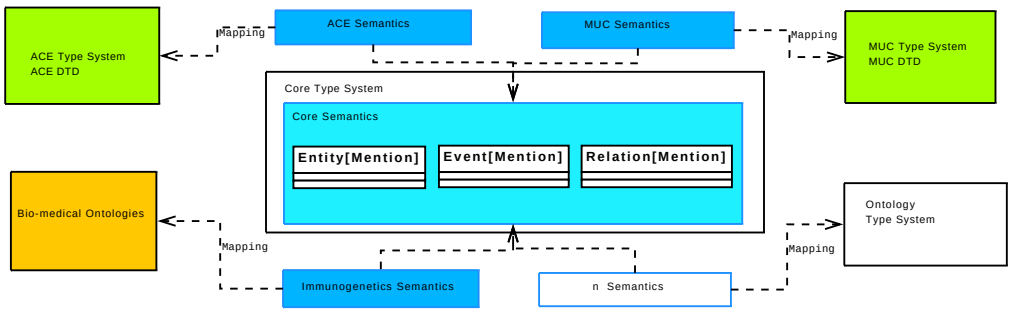


Figure 2: Extensions of the *Semantics* Layer

scrutiny. The feature *specificType* describes the relation in more depth, supplying a subtype (usually *is-a*) of the original relation, e.g., `locatedIn` is a *specificType* of a `Part-of` relation.

`EventMention` stands for a special kind of semantic relations between entities (Doddington et al., 2004), namely those representing n-ary semantic relations. The `EventMention` type is supplied with seven additional features (see Figure 1). The *begin* and *end* features refer to the text region that marks the event occurrence (e.g., word). *arguments* is a feature which contains an array of *n* `ArgumentMention`(s). An `ArgumentMention` refers through the feature *mention* to the `EntityMention` in the text as an event participant. The feature *role* assigns a certain role to the argument in the event under scrutiny. *Polarity* is needed to distinguish between true and negated events. The features *tense* and *modality* characterize the time the event/relation occurs (e.g., future, past) and the information about how certain this event/relation is (e.g., asserted, hypothetical), respectively.

The introduced types constitute a platform ready for domain-specific extensions. In the following, we show how to envelop already established semantic resources within such an annotation type system.

5 Domain-Specific Extensions of the *Semantics* Layer

We extend here the core semantics type system with domain-specific semantics for the newswire and the biomedical domains (see Figure 2). For the newswire area we selected the ACE taxonomy. In the biomedical domain we cover parts of the areas of *immunogenetics* and *regulation of gene expression*.[3]

5.1 Newswire Domain

The general objective of the *Automatic Content Program* (ACE) is to develop information extraction technology, with focus on entity, relation and event recognition (Doddington et al., 2004). For this task, a two-level taxonomy is provided to distinguish between types and subtypes in the annotated newspaper data. The entity taxonomy contains types such as *Person*, *Organization*, *Location*; subtypes of *Person* are e.g., *Group*, *Individual*. The relation taxonomy contains types such as *Physical*, *Part-Whole*; subtypes of *Physical* are e.g., *Located*, *Near*. The event taxonomy contains types such as *Life*, *Contact*; subtypes of *Life* are e.g., *Be-Born*, *Marry*.

We represent the ACE taxonomy as a CAS type system by creating CAS types from ACE types (see Figure 3). The ACE *Semantics* type system is an extension of the core type system just introduced. All ACE CAS types extend the core semantic types `EntityMention`, `RelationMention` or `EventMention`, accordingly (e.g., Person(`Person`) extends `EntityMention`, Part-Whole(`Part-Whole`) extends `RelationMention`, while Conflict(`Conflict`) extends `EventMention`).

The ACE subtypes can be represented in CAS using the feature *specificType* that describes the semantic type in a more detailed way (e.g., *specificType* of `Person` may be `Individual`). In this

[3]The field of *immunogenetics* is the domain focus in the STEMNET project (`http://www.stemnet.de`). The field of *regulation of gene expression* is one of the domain focus areas in the BOOTSTREP project (`http://www.bootstrep.org`).

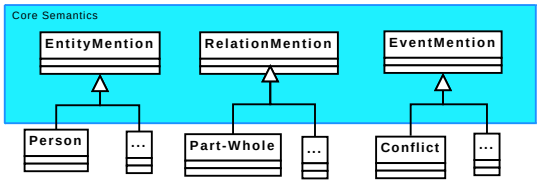


Figure 3: Extensions of the *Semantics* Layer for the ACE Semantics

way, we represent the two-level ACE taxonomy as ACE types using the *specificType* feature.

In order to dock the complete ACE corpus to the UIMA Framework, we converted the ACE XML Document Type Definition (DTD) to the CAS type system definition by creating CAS types from DTD elements independent of the core type system. Every element in the DTD becomes a type in the CAS. The features of the types are the children and attributes of the corresponding element in the DTD. The complete ACE annotation was converted this way to the CAS representation.

Both type systems are connected by a mapping file which allows for the conversion of the CAS ACE annotation based on the ACE type system to the ACE *Semantics*. The benefit of the creation of the ACE *Semantics* type system is the compatibility to the core type system. The ACE semantic types extend the core semantic types and can be analyzed by tools without any specifications for the original ACE annotation schemata (i.e., the ACE DTD).

5.2 Biomedical Domain

In the biomedical domain available ontological resources play a considerable role in the establishment of the annotation terminology.[4] When we move to the automatic annotation of biomedical documents within an NLP framework, we have to cope with the interrelations between the terminology as available in an ontology and the type system for semantic annotation.

The ontology languages being currently used (e.g., OWL (Patel-Schneider et al., 2002)) are richer in their expressivity (e.g., multiple inheritance) than a graph-based annotation model implemented in CAS. Therefore, the aim is certainly not to re-implement ontologies as CAS type systems but rather to provide a linkage of CAS annotations to the original ontological resources via a clean interface definition.

For the biological application, we thus designed a type system which covers a substantial portion of e.g.., the area of the *immunogenetics*. As ontological resources we used, e.g., Gene Ontology, IMR Ontology and Cell Ontology.[5]

CAS annotations are linked to the original ontological source by the feature *specificType* that provides the ontology source and entry identifier. The information about this mapping between CAS types and their ontological entries is indexed in mapping files which can be used by NLP components requiring the information about the ontological representation of CAS types.

In the following we demonstrate the extended type system in action within the UIMA framework.

6 Use Cases

The UIMA framework not only provides a formal specification layer based on UML,[6] but also comes with a run-time environment for the interpretation of the type system specifications. This turned out to be useful for implementing systems based on UIMA specifications some of which we introduce below.

6.1 NLP Pipelines

For the purposes of information extraction and retrieval in the area of *immunogenetics*, we set up the STEMNET pipeline (Hahn et al., 2007b) which contains a variety of syntactic and semantic components and a database indexer. For the pipeline configuration we took the core type system and the *immunogenetics* semantics type system. We represent all results of the NLP processing in the CAS format and maintain the linkage to the original source ontologies. This allows for example for the indexing

[4] http://obo.sourceforge.net

[5] http://obofoundry.org

[6] http://www.uml.org

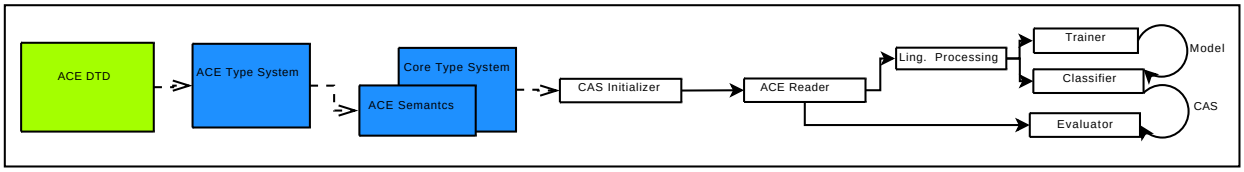


Figure 4: ACE Pipeline, Supervised Approach

of annotation results with ontological entries for information retrieval applications. A further benefit would be the conversion of CAS annotation results into a more expressive representation format (RDF[7] graphs) in order to perform reasoning over results as shown by Welty and Murdock (2006).

6.2 Support for supervised approaches

Previous research provides ample evidence that (ML-based) supervised approaches are very effective for named entity recognition and relation extraction (e.g., Settles (2004) or Zhang et al. (2006)). In order to support the generation of training data for supervised approaches, semantically annotated corpora (e.g., ACE) have been built. Usually, semantically annotated corpora lack additional annotation layers. As the supervised approaches in semantic analysis profit from the linguistic pre-processing annotation (i.e., segmentation, morpho-syntactic and syntactic annotation), corpora should be automatically extended relative to annotation layers not already accounted for.

Since the additional annotations are usually not provided in the format of the original corpora, one has to switch between various formats in order to retrieve the annotations of interest. This decreases the performance of the analysis and can be error-prone.

Because we extended a core type system with domain-specific ACE semantics, we can conveniently represent all results of linguistic analysis together with the ACE semantics as combined in the CAS. The ACE pipeline (see Figure 4) reveals how to deal with various linguistic annotations for the purposes of supervised approaches. The pipeline provides an ACE *trainer*, an ACE *classifier* and an ACE *evaluator*. It starts with data segmentation in the training and test data. Data processing starts with the ACE READER which converts the original ACE data to the CAS representation format. Next, the data is processed by linguistic analysis engines (i.e, segmentation, POS tagging, chunking, parsing). After the data is pre-processed, the training CAS data is sent to the *trainer* which can extract features from the CAS and produces a classification model. The model is then used by the *classifier* that processes the CAS test data and writes the results to the CAS. The *evaluator* compares the results of the *classifier* with the gold standard.

6.3 Interoperability and Visualization

All CAS annotations may be serialized in the XMI format within the UIMA framework. XMI, the XML Metada Interchange format (stand-off annotation), is an OMG[8] standard for the XML representation of object graphs. The CAS annotations can be visualized within the XCAS Visualizer in the UIMA framework. The visualization provides a user-friendly look inside the data.

7 Conclusion and Future work

In this paper, we proposed an integration of various semantic annotation languages, for general newspaper language as well as for (parts of) the sublanguage of the life science, into the coherent framework of an annotation type system. While previous efforts at lower levels of linguistic analysis, up until the syntactic level, are characterized by a significant degree of consensus in what concerns the basic vocabulary of such specification layers, the semantic layer is much harder to deal with. In essence, the semantics layer might contain a multitude of entities similar to entries in a lexicon or concepts in an ontology. So care has to be taken not to duplicate information contained in these resources at the semantic annotation layer.

[7] http://www.w3.org/RDF/

[8] http://www.omg.org

Here, type hierarchies might be of help for making higher levels of abstraction explicit. Still, they do not provide any heuristic guidance at which level of specificity one has to stop in terms of increasing the level of detail for semantic annotation.

We here defined a clean interface between generic semantic annotation types and the highly elaborated granularity of the conceptual structure of ontologies (or lexical semantic specifications kept in a lexicon). This way, we avoid the redundant duplication of lexical and conceptual knowledge and keep the number of linguistically relevant semantic entities used for semantic annotation at a manageable size.

Acknowledgments. This research was funded by the EC's 6th Framework Programme (4th call) within the BOOTStrep project under grant FP6-028099, and by the German Ministry of Education and Research (BMBF) via its e-Science initiative within the StemNet project (funding code: 01DS001A to 1C).

References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1/2):23–60.

Paul Buitelaar, Thierry Declerck, Bogdan Sacaleanu, Špela Vintar, Diana Raileanu, and Claudia Crispi. 2003. A multi-layered, XML-based approach to the integration of linguistic and semantic annotations. In *Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML-03)*. Budapest, Hungary, April, 2003.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program: Tasks, data, & evaluation. In *LREC 2004 – Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 837–840. Lisbon, Portugal, 26-28 May 2004.

David Ferrucci and Adam Lally. 2004. UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Thilo Götz and Oliver Suhre. 2004. Design and implementation of the UIMA Common Analysis System. *IBM Systems Journal*, 43(3):476–489.

Udo Hahn, Ekaterina Buyko, Katrin Tomanek, Scott Piao, John McNaught, Yoshimasa Tsuruoka, and Sophia Ananiadou. 2007a. An annotation type system for a data-driven NLP pipeline. In *The LAW at ACL 2007 – Proceedings of the Linguistic Annotation Workshop*, pages 33–40. Prague, Czech Republic, June 28-29, 2007.

Udo Hahn, Joachim Wermter, David S. DeLuca, Peter Horn, Rainer Blasczyk, Michael Poprat, and Asad Bajwa. 2007b. STEMNET: An evolving service for knowledge networking in the life sciences. In *Proceedings of the German e-Science Conference 2007*. Baden-Baden, Germany, May 2-4, 2007.

Nancy Ide and Keith Suderman. 2007. GRAF: A graph-based format for linguistic annotations. In *The LAW at ACL 2007 – Proceedings of the Linguistic Annotation Workshop*, pages 1–8. Prague, Czech Republic, June 28-29, 2007.

Nancy Ide, Laurent Romary, and Eric de la Clergerie. 2003. International standard for a linguistic annotation framework. In *Proceedings of the HLT-NAACL 2003 Workshop on 'Software Engineering and Architecture of Language Technology Systems'*, pages 25–30. Edmonton, Canada, May 31, 2003.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The PENN TREEBANK. *Computational Linguistics*, 19(2):313–330.

Adam Meyers. 2006. Annotation compatibility working group report. *Proceedings of the COLING-ACL 2006 Workshop on 'Frontiers in Linguistically Annotated Corpora 2006'*, pages 38–53. Sydney, Australia, 22 July 2006.

Tomoko Ohta, Y. Tateisi, and J. Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. *HLT 2002 – Proceedings of 2nd International Conference on Human Language Technology Research*, pages 82–86. San Diego, CA, USA, March 24-27, 2002.

Peter F. Patel-Schneider, Pat Hayes, Ian Horrocks, and Frank van Harmelen. 2002. OWL Web Ontology Language; Semantics and Abstract Syntax. `[http://www.w3.org/TR/owl-semantics/]`.

Scott Piao, Ekaterina Buyko, Yoshimasa Tsuruoka, Katrin Tomanek, Jin-Dong Kim, John McNaught, Udo Hahn, Jian Su, and Sophia Ananiadou. 2007. Bootstrep annotation scheme: Encoding information for text mining. In *Corpus Linguistics 2007 – Proceedings of the 4th Corpus Linguistics Conference*. Birmingham, England, U.K., July 27-30, 2007.

James Pustejovsky, José Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. TIMEML: Robust specification of event and temporal expressions in text. In Mark Maybury, editor, *New Directions in Question Answering*, pages 28–34.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, and M. Lazo. 2003b. The TIMEBANK corpus. *Proceedings of the Corpus Linguistics 2003 Conference*, pages 647–656. Lancaster, U.K., 28-31 March 2003.

Ulrich Schäfer. 2006. Middleware for creating and combining multi-dimensional NLP markup. In *Proceedings of the 5th EACL-2006 Workshop on NLP and XML (NLPXML-2006): Multi-Dimensional Markup in Natural Language Processing*, pages 81–84. Trento, Italy, April 4, 2006.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. *JNLPBA – Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 107–110. Geneva, Switzerland, August 28-29, 2004.

Christopher A. Welty and J.W̃illiam Murdock. 2006. Towards knowledge acquisition from information extraction. *The Semantic Web – ISWC 2006. Proceedings of the 5th International Semantic Web Conference*, pages 709–722. Athens, GA, USA, November 5-9, 2006.

Min Zhang, Jie Zhang, Jian Su, and GouDong Zhou. 2006. A composite kernel to extract relations between entities with both flat and structured features. In *COLING'ACL 2006 – Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 825–832. Sydney, Australia, July 17-21, 2006.

Approaches towards a "Lexical Web": the role of Interoperability

Nicoletta Calzolari
Istituto di Linguistica Computazionale del
CNR / 56100 Pisa, Italy
glottolo@ilc.cnr.it

Abstract

After highlighting some of the major dimensions that are relevant for Language Resources (LR) and contribute to their infrastructural role, I underline some priority areas of concern today with respect to implementing an open Language Infrastructure, and specifically what we could call a "Lexical Web". My objective is to show that it is imperative to define an underlying global strategy behind the set of initiatives which are/can be launched in Europe and world-wide, and that it is necessary an all-embracing vision and a cooperation among different communities to achieve more coherent and useful results. I end up mentioning two new European initiatives that go in this direction and promise to be influential in shaping the future of the LR area.

1 Language Resources: major dimensions

Only in the '90s LRs started to be considered as the necessary platform on which technologies and applications are built, a recognition which is nowadays widely accepted for the takeoff of our field. The following types of initiatives were then considered the major building blocks to set up a LR infrastructure (Calzolari and Zampolli, 1999):

i) *Standards for LRs*: the concept of reusability – directly related to the importance of "large scale" LRs within the increasingly dominant data-driven approach – has contributed significantly to the structure of many R&D efforts, such as EAGLES, ISLE, the recent LIRICS (e-Content), the ISO-TC37/SC4 committee.
ii) *LR construction*: projects such as WordNet, PAROLE, SIMPLE, LC-Star, EuroWordNet.
iii) *LR distribution*: LDC (Linguistic Data Consortium) in US, ELRA (European Language Resources Association) in Europe.

Other dimensions were soon added as necessary complement to achieve the required robustness and data coverage and to assess results obtained with current methodologies and techniques, i.e.:

iv) *Automatic acquisition of LRs* or of linguistic information: projects such as ACQUILEX, SPARKLE, ECRAN.
v) *Use of LRs for evaluation* campaigns, such as MUC, TREC, CLEF, Senseval, ACE.

1.1 Success of the Field

The very large body of initiatives of the last two decades (Calzolari, 1998 for an overview) was instrumental for the formation of a "LR community", and gave rise to a set of international initiatives of a global nature, encompassing many various perspectives on LRs or dealing with policy and meta-level issues related to LRs, such as:

- The Thematic Network ENABLER, grouping European National projects on LRs;
- The LREC Conference (about 900 participants in Lisbon-2004 and Genova-2006);
- The Asian Federation of Natural Language Processing (AFNLP);
- Bodies such as COCOSDA (International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques) and WRITE (Written Resources Infrastructure, Technology and Evaluation);
- The new journal *Language Resources and Evaluation* (Ide and Calzolari, 2005).

Not to mention the ever-increasing role of LRs in statistical and empirical methods, and the growing industrial interest in using LRs and standards, specially for multilingual applications.

The flourishing of international projects and activities contributed to substantially advance knowledge and capability of how to represent, create, acquire, access, tune, maintain, standardise, etc. large lexical and textual repositories. There are today countless initiatives in the LR field, but we must admit that they are somehow scattered, opportunistic, often unconnected, with no real ability to build on each other and to form a unified space of LRs. We thus recognise that the LR infrastructure is still a virtual one. There is no real global coordination of efforts, and no body able to create the needed synergies among the various initiatives.

On the other side, the success itself of the field, its vitality and richness, coupled with the lack of coordination and of strategic thinking about future orientations, show that it is time to reflect again on the field as a whole, and ask ourselves which are/will be the major driving forces of today and of tomorrow to give the field the necessary cohesion.

1.2 Need of a Change

The wealth of LRs, in comparison with few years ago, but coupled with the shortage, even now, of a) new types of LRs, b) multilingual LRs, c) LRs of much larger size, d) LRs with richer annotations, and so on, points towards the need to consider whether those mentioned above are still the major driving forces. Which new building blocks do emerge today? I believe that those dimensions are still relevant, even if with an obvious evolution. Emerging pillars in current HLT are:

i) *Interoperability*, and even more *content interoperability*: language is the key mediator to access content, knowledge, ontologies;
ii) *Collaborative creation and management of LRs*, even on the model of wiki initiatives;
iii) *Sharing of LRs*, as a new dimension of the distribution notion;
iv) *Dynamic LRs*, able to auto-enrich themselves;

and finally the more comprehensive notion of:

v) *Distributed architectures and infrastructures for LRs*, encompassing and exploiting the realisation of the previous notions.

I will mention in the last section two new European initiatives where such notions will play a prominent role, could be at the basis of a new paradigm for LRs and language technology (LT) and influence the setting up of a "real" infrastructure.

2 Some Tendencies and Driving Forces in the Lexical Domain

Mixing considerations on what is needed for a broad language infrastructure and for a "lexical web" – undoubtedly a key part of it –, I touch here issues relevant to establishing a lexical web. I do that by pointing at research activities carried out at ILC in Pisa[1] showing a variety of approaches to lexical resources, involving: i) procedures for linking and integrating existing lexicons, ii) standardisation, iii) rel ation between lexical and terminological or ontological resources, iv) "ontologisation" of lexicons, v) architectures for managing, merging, integrating lexical resources.

2.1 Integration/Unification of Existing Lexicons

The market is increasingly calling for new types of lexicons that can be built rapidly – tailored to specific requirements – possibly by combining certain types of information from available lexicons while discarding others. This need could be satisfied exploiting the richness of existing lexicons, aiming at attaining their integration or virtual unification.

ELRA Unified Lexicon. An initiative in this direction, the *Unified Lexicon* project, has been carried out at ELRA by its Production Committee (Monachini et al, 2006). This experiment consisted in linking the LC-Star and PAROLE lexicons to set up a methodology to connect Spoken and Written LRs, thus establishing new models of LR distribution. In the envisaged scenario the same lexicons may be made available to different users, who can select different portions of the same lexicon or combine information coming from different lexicons. In this scenario lexical resources can be shared, are reusable and openly customisable, instead of being static and closed.

Linking ItalWordNet and SIMPLE Semantic Lexicons. The two largest and extensively encoded Italian lexicons, ItalWordNet (IWN) and PAROLE-SIMPLE-CLIPS (PSC), although developed according to two different lexical models, present many compatible aspects. Linking – and eventually merging – these lexicons in a common representation framework means to offer the end-user more exhaustive lexical information combining poten-

[1] Many passages in this section are taken from various papers, listed in the References, of ILC colleagues.

tialities and outstanding features offered by the two lexical models (Roventini et al, 2007). Not only reciprocal enhancements are obtained, but also a validation of the two resources. Their semantic integration is all the more desirable considering their multilingual vocation: IWN is linked to wordnets for many other languages, and PSC shares with 11 European lexicons theoretical model, representation language, building methodology and a core of entries.

Mapping the Ontologies and the Lexicons. Due to a different organisational structure of the two ontology-based lexicons, the linking process involves elements having a different status, i.e. autonomous semantic units in PSC and synsets in IWN. Mapping is performed on a semantic type-driven basis: comparing their ontological framework and establishing correspondences between the conceptual classes of both ontologies, with a view to further matching their respective instances, using also 'isa' relations and semantic features. The result of the first phase, linking concrete entities, sounds promising since 72.32% of the word-senses have been successfully linked.

The linking process makes it possible to enrich each resource by complementary information types peculiar to the other's theoretical model. In IWN, the richness of sense distinctions and the consistency of hierarchical links are remarkable. SIMPLE focuses on richly describing the meaning and semantic context of a word and on linking its syntactic and semantic representation, crucial for most NLP applications. Moreover, the mapping lets inconsistencies emerge, allowing to amend them. The linking process implies a de facto reciprocal assessment of both coverage and accuracy, particularly relevant to hand-built lexicons.

Differences regarding the nature of linking units, the granularity of sense distinction and the ontological typing are complex issues that are also being addressed during the linking process.

2.2 Interoperability: at the Heart of the Field

We have made big steps forward with respect to interoperability. Work started in EAGLES and ISLE (www.ilc.cnr.it/EAGLES96/isle/) (Calzolari et al, 2003) is being recently consolidated in true international ISO standards.

ISO. The Working Group ISO TC37/SC4/WG4 dedicated to NLP lexicons is in charge of defining lexical standards. The result is the LMF (Lexical Markup Framework) standard (Francopoulo et al, 2006). To cope with the challenge that actual lexicons differ very much both in complexity and in type of encoded information, a modular organization was adopted. As a consequence, LMF (http://lirics.loria.fr/documents.html) is made up of a core model, a sort of simple skeleton, and various semi-independent packages of notions, used for the various linguistic layers that make up a lexicon.

Lexical specifications are split in separate object types: LMF defines the lexical structure and is kept simple, while the huge amount of attributes (e.g. Part-of-Speech) are recorded in a data category registry where the peculiarities of languages and linguistic schools can be recorded. This registry, common to all TC37/SC4 standards, guarantees interoperability between lexicon and corpus annotation. An XML DTD is based on the UML modelling. Moreover, an OWL format has been defined that can be smoothly integrated into Semantic Web applications.

NEDO. While EAGLES and ISLE dealt with European languages, the Japanese NEDO project (Tokunaga et al, 2006), that develops international standards for Semantic Web applications, is specifically geared to Asian languages: Chinese, Japanese, Thai. It applies and refines ISO standards so that they are adapted to Asian languages.

But true content interoperability is still far away. We may have solved the issue of formats, of inventories of linguistic categories for the various linguistic layers, but have not solved the problem of relating senses, that only would allow automatic integration of semantic resources. This is a challenge for the next years, and a prerequisite for both a true Lexical Web and a credible Semantic Web.

2.3 Lexicons vs. Terminologies: a Continuum

Due to the strategic relevance of the biomedical field, intensive research is being carried out worldwide to develop LTs to access its large body of literature and extract knowledge from it. Access to and interoperability of biological databases, however, is still hampered by lack of uniformity and harmonisation of both formats and information encoded. A current demand in bioinformatics is to construct a comprehensive and incremental resource which integrates bio-terms encoded in existing different databases. A challenge is to encode all relevant properties of bio-terms according to the most accredited standards for the representation of

lexical, terminological and conceptual information.

BioLexicon. Working in the bio-domain in the European BOOTStrep project[2], we assume that the linguistic side of terminologies is partially informed by the knowledge of the domain and we claim that semantic relations, especially those accounting for the syntagmatic relations of words in context, are crucial for the representation of this kind of information. We also argue (Monachini et al, 2007) that a privileged representational device for encoding these relations is the set of Qualia Relations, as encoded in the SIMPLE general lexicon. These assumptions are made operational in the design of the *BioLexicon*: building a comprehensive terminological resource for the biomedical domain – with morphological, syntactic, semantic descriptions of the terms – which adheres to lexical and ontological standards and links concepts to lexical items is a huge scientific challenge. The BioLexicon (Quochi et al, 2007) is a large-scale resource that combines terminological data coming from bio-databases (mostly UniProt, Swiss-Prot, ChEBI, BioThesaurus and NCBI taxonomy) enriched with lexical information extracted from texts. The lexicon model is designed so as to integrate both typical information provided by domain ontologies and linguistic information available in open-domain computational lexicons: terms and variants are encoded with their semantic information as well as with typical linguistic information such as Part-of-Speech, subcategorisation frames and qualia relations that can be further augmented and tuned to cope with domain specific semantic information.

The model – flexible enough to adapt to different application needs, e.g. text-mining, information extraction, information retrieval, multilingual access – builds on previous experience in the standardisation and construction of lexical resources. Both the conceptual model and its physical implementation are tailored to the automatic population of the resource, independently of the various native data formats. The DB is modular and can automatically upload new data and provide (XML) outputs by means of web services. An *XML interchange format* (XIF) has been designed with the purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems, therefore allowing for a standardisation of the data extracted from the different terminological resources and from texts.

The goal is to propose a standard for the representation of lexicons in the bio-domain, which could eventually be also interoperable with other domain lexicons. For this reason the ISO LMF was chosen as the reference meta-model and the ISO Data Categories as the main building blocks for the representation of the entries. A reusable BioLexicon with sophisticated linguistic information, linked to a bio-ontology, should enable the bio-informatics community to develop information extraction tools of higher quality.

2.4 Lexicons and Ontologies : a Dilemma

Ontologies are recognised as an important component in NLP systems that deal with the semantic level of language (Huang et al, to appear). Most semantic lexical resources (e.g. WordNet, CYC, SIMPLE), have in common the presence of an ontology as a core module. Besides, there is a lot of research in progress on applying ontologies to semantic NLP. The fact that OWL is the ontology language for the Semantic Web and that it provides a formal semantic representation as well as reasoning capabilities has encouraged the NLP community to convert existing resources to this language.

RDF/OWL representation of WordNet. WordNet has recently received a growing attention by the Semantic Web community. Within W3C, WordNet has been translated (Van Assen et al, 2006) in the standard semantic languages RDF/ OWL, which can describe collections of resources on the Web and are convenient data models to represent highly interconnected information and their semantic relations. Moreover, the RDF/OWL representation of WordNet is easily extensible, allows for interoperability and makes no assumptions about particular applications. The availability of WordNet Web Services can be an important step for its integration and effective use into the Semantic Web, and for future multilingual semantic interoperability in the Web (Marchetti et al, 2006).

"Ontologisation" of lexicons. A new initiative at ILC (Toral and Monachini, 2007) is the conversion into OWL of the ontology of SIMPLE, the lexico-semantic resource based on the Generative Lexicon (GL). The elements of SIMPLE modelled in OWL are those of the original ontology, i.e. se-

[2] BOOTStrep (Bootstrapping Of Ontologies and Terminologies STrategic Project) is an IST European project under the 6th Framework Programme (www.bootstrep.eu).

mantic types, qualia relations, semantic features. A challenge in the ontology design is that its nodes are not only defined by their formal dimension (taxonomic hierarchy), but also by the GL qualia dimensions: constitutive, telic, agentive. The OWL ontology is also enriched in a bottom-up approach that extracts further semantic information (e.g. selected constraints on relations and features extracted from the lexicon) by exploring the word-senses that belong to each semantic type and by using the qualia structure as a generative device.

This research aims at the representation of a lexicon based in the GL theory into the Semantic Web ontology language, with reasoning capabilities interfaced to a lexicon. This allows the ontology to be processed and checked by standard reasoners. This is useful for Semantic Web applications, semantic NLP tasks, and for enhancing the quality of the lexicon by validating it (through reasoning one can look for inconsistencies). The ontology is also a key element of a broader forthcoming research aimed at automatic lexico-semantic-driven text mining and knowledge acquisition procedures, which, in their turn, have the goal of gathering knowledge to enrich the lexicon, thus creating a virtuous circle between lexicon/ontology and corpus-based information acquisition.

2.5 Architectures for Integration of Lexicons

Enhancing the development of multilingual lexicons is of foremost importance for intercultural collaboration (Bertagna et al, 2007), as they are the cornerstone of several multilingual applications. Nevertheless, large-scale multilingual lexicons are not yet as widely available as needed. A new trend tries to exploit the richness of existing lexicons, in addition to creating new ones. At the same time, as the history of the web teaches, it would be a mistake to create a central repository of all the shared lexicons, while distribution of resources is a crucial concept. A solution emerging in the LR community consists in moving towards distributed language services, based on open content interoperability standards, and made accessible to users via web-service technologies. There is another deeper argument in favour of distributed lexical resources: LRs are inherently distributed because of the diversity of languages over the world. It is natural that LRs are developed and maintained in their native environment. It is not possible to describe the current state of a language, evolving over time, away from where it is spoken.

Web services for LRs or LRs as web services. Having lexicons available as web services would allow to create new resources on the basis of existing ones, to exchange and integrate information across repositories and to compose new services on demand: an approach towards the development of an infrastructure built on top of the Internet in the form of distributed language services is presented in Ishida (2006). This new type of LRs can still be stored locally, but their maintenance and exploitation can be a matter of agents choreographed to act over them. Admittedly, this is a long-term scenario requiring the contribution of many actors and initiatives (among which we mention standardisation, distribution, international cooperation). A first prerequisite for this scenario to take place is to ensure true interoperability among lexicons, a goal that would be now mature for many aspects. Although the paradigm of distributed and interoperable lexicons has largely been discussed and invoked, little has been made for its practical realisation. Some initial steps to design frameworks enabling interlexica access, search, integration, operability are: the Lexus tool (Kemps-Snijders et al, 2006), based on LMF, managing the exchange of data among large-scale lexical resources, and SHAWEL (Gulrajani and Harrison, 2002), tailored to the collaborative creation of lexicons for endangered language. However, the impression is that little has been made towards the development of new methods, techniques and tools for attaining a real interoperability among lexical resources.

LeXFlow. The design of an architecture able to turn into reality the vision of shared and distributed lexical repositories is a very challenging task. To meet these needs, we have designed and built a distributed architecture, *LeXFlow*, enabling a rapid prototyping of cooperative applications for integrating lexical resources (Soria et al, 2006). It is based on a web-service architecture, fostering integration and interoperability of computational lexicons, focusing on mutual linking and cross-lingual enrichment of distributed monolingual lexicons. As case–studies, we have chosen to work with:

i) two Italian lexicons based on different models, SIMPLE and ItalWordNet, and
ii) two lexicons belonging to the WordNet family, ItalWordNet and the Chinese Sinica BOW.

These represent different opportunities of adopting a bottom–up approach to exploring interopera-

bility for lexicon augmentation and mutual enrichment of lexical resources, either i) in a cross-model or ii) in a cross-lingual enrichment/ fertilisation of monolingual lexicons.

Multilingual WordNet Service. This module is responsible for the automatic cross-lingual fertilisation of lexicons with a wordnet-like structure. Put it very simply, the idea behind it is that a monolingual WordNet can be enriched by accessing the semantic information encoded in corresponding entries of other monolingual WordNets. The various WordNet-lexicons reside over distributed servers and can be queried through web service interfaces. The entire mechanism is based on the exploitation of the Interlingual Index (ILI). The proposal to make distributed WordNets interoperable allows applications such as:

– *Enriching existing resources*. Information is not complete in any WordNet: by making WordNets interoperable we can bootstrap semantic relations and other information from other WordNets.
– *Creation of new resources*. Multilingual lexicons can be bootstrapped by linking different language WordNets through the ILI.
– *Validation of resources*. Semantic relations and synset assignments can be validated if reinforced by data coming from other WordNets.

This work can be a prototype of a web application to support the *Global WordNet Grid* initiative (www.globalwordnet.org/) (Fellbaum and Vossen, 2007), whose success depends on whether there will be tools to access and manipulate the rich internal semantic structure of distributed multilingual WordNets. *LeXFlow* offers such a tool, providing interoperable web-services to access distributed WordNets on the grid. This allows to exploit in a cross-lingual framework the wealth of monolingual lexical information built in the last decade. As an example of use, a multilingual query given in Italian but intended for querying English, Chinese, French, German, and Czech texts, can be sent to 5 different nodes on the Grid for query expansion, as well as performing the query itself. This way, language-specific query techniques can be applied in parallel to achieve results that can be then integrated. As multilingualism clearly becomes one of the major challenges of the future of web-based knowledge engineering, WordNet emerges as a leading candidate for a shared platform, representing a simple and clear lexical knowledge model for different languages. This is true even if it has to be recognised that the WordNet model is lacking some important semantic information (like a way to represent semantic predicates). In LeXFlow we presuppose a de-facto standard, i.e. a shared and conventionalised architecture. Since the WordNet framework is both conventionalised and widely followed, our system is able to rely on it without resorting to a more substantial and comprehensive standard. In the case, however, of integration of lexicons with different underlying linguistic models, the availability of MILE (Calzolari et al, 2003), now of LMF, is an essential prerequisite.

From a more general viewpoint, we must note that the realisation of the new vision of distributed and interoperable LRs is strictly intertwined with at least two prerequisites. On the one side, LRs need to be available over the web; on the other, the LR community will have to reconsider current distribution policies, and investigate the possibility of developing an "Open Source" concept for LRs.

UIMA. Finally, we have started an initiative, at ILC, to integrate both various LRs (lexicons, ontologies, corpora, etc.) and different NLP tools into a common framework of shared and distributed resources, the IBM UIMA middleware (Ferrucci and Lally, 2004). As case study, a first prototype for a UIMA Type System has been built to manage TimeML categories and integrate an Italian Treebank and the SIMPLE lexicon (Caselli et al, 2007). Both a web interface for human access and a series of web services for machine use are being developed. This research intends to contribute both to a UIMA type systems standardisation and to a common framework for resource and tool sharing and interoperability definition. This initiative is linked with the NICT Language Grid project (Ishida, 2006), from which our prototype inherits the service ontology environment.

3 First steps for a LR Infrastructure

Finally, new conditions are emerging, in Europe, that could turn what is so far a virtual LR infrastructure into a real one (Calzolari, 2007). This tendency is helped not only by new technical conditions, but also by the recognition that any organisation has limited resources, and will never be able to create all the necessary infrastructural resources – in adequate quality – as needed. These may instead be spread across several organisational

units.

Sensitivity of LRs: political, economic, social, strategic factors. Behind the notion of "distributed" resources there are also political (very sensitive) factors, behind resources that can be "shared/ and reused" economic factors. Moreover, many today start bringing into focus also the social value of a common infrastructure, and strongly advocate – contrary to current practice – the benefits of open access (vs. the social costs of restricted access). In addition to its scientific implications, the large intellectual, cultural, economic movement behind LRs entails "strategic" thinking, and urges to reflect on field of LRs from a very broad angle. It is perceived as essential to define a general plan for research, development and cooperation in the LR area, to avoid duplication of efforts and provide for a systematic distribution and sharing of knowledge. To ensure reusability, the creation of standards is still the first priority. Another tenet is the recognition of the need of a global strategic vision, encompassing different types of – and different methodologies of building – LRs, for an articulated and coherent development of this field.

Two new European initiatives are linked to these ideas.

3.1 CLARIN

CLARIN (*Common Language Resource and Technology Infrastructure*) (http://www.mpi.nl/clarin/) is an ESFRI project whose mission is to create an infrastructure that makes LRs and LTs easily usable to scholars of all disciplines, in particular of the humanities and social sciences, to prepare an eScience scenario. The purpose is to offer persistent and secure services and provide easy access to LRs and LTs. CLARIN proposes to make this vision a reality: the user will have access to repositories of data with standardised descriptions, processing tools ready to operate on standardised data, and guidance from distributed knowledge centres. All this will be available on the web using a service oriented architecture based on secure grid technologies. CLARIN will turn existing, fragmented LRs and LTs into accessible, stable services that any user can share, adapt and repurpose, building upon the rich history of European and national initiatives. The preparatory phase aims at bringing the project to the required level of legal, organisational and financial maturity. This necessitates an approach along various dimensions in order to pave the way for implementation. Infrastructure building is a time-consuming activity and only robustness and persistency of the offered solutions will convince researchers and users.

3.2 FLaReNet

International cooperation and re-creation of the LR community are among the most important drivers for a coherent evolution of the LR area in the next years. The Thematic Network *FLaReNet (Fostering Language Resources Network),* proposed in the context of an eContent*plus* call, will act as a European forum to facilitate interaction among LR stakeholders. Its structure considers that LRs present various dimensions and must be approached from many angles: technical, but also organisational, economic, legal, political, addressing also multicultural and multilingual aspects, essential when facing access and use of digital content in today's Europe. FLaReNet, organised into working groups focusing on specific objectives, will bring together leading experts (academic and industrial) to ensure, in cooperation with CLARIN, coherence of LR-related efforts in Europe. FLaReNet will consolidate existing knowledge, presenting it analytically and visibly, and will contribute to structuring the area of LRs of the future by discussing new strategies to: convert existing and experimental technologies related to LRs into useful economic and societal benefits; integrate so far partial solutions into broader infrastructures; consolidate areas mature enough for recommendation of best practices; anticipate the needs of new types of LRs.

The outcome of FLaReNet will be of a directive nature, to shape the future of the LR area, and help the EC, and national funding agencies, to identify the priority areas of LRs that need public funding to develop and improve. A blueprint of actions will give input to policy development both at EU and national level for identifying new language policies that support linguistic diversity in Europe, in combination with strengthening the language product market and introducing innovative services, especially for less technologically advanced languages.

References

Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C., Hsieh, S., Marchetti, A., Tesconi, M., Fostering Intercultural Collaboration: a Web Service

Architecture for Cross-Fertilization of Distributed Wordnets. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of the First International Workshop on Intercultural Collaboration (IWIC 2007)*, Kyoto, pp. 185-198. Also in: LNCS, vol. 4568, pp. 146-158. Springer, 2007.

Calzolari, N., An overview of Written Language Resources in Europe: a few reflections, facts, and a vision. In *Proceedings of the First LREC*, pp. 217-224. Granada, 1998.

Calzolari N., Towards a new generation of Language Resources in the Semantic Web vision. In Ahmad, K., Brewster, C., Stevenson, M. (eds.), *Words and Intelligence II: Essays in honour of Yorick Wilks*, pp. 63-105. Springer, 2007.

Calzolari, N., Bertagna, F., Lenci, A., Monachini, M. (eds.), *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry).* ISLE, Pisa, 194 pp., 2003.

Calzolari, N., Zampolli, A., Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure. In *MT Summit Proceedings*, pp. 358-365. Singapore, 1999.

Caselli, T., Prodanof, I., Ruimy, N., Calzolari, N., Mapping SIMPLE and TimeML: improving event identification and classification using a semantic lexicon. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Fellbaum, C., Vossen. P., Connecting the Universal to the Specific: Towards the Global Grid. In Ishida, T., Fussell, S. R., Vossen, P. (eds.) *Proceedings of IWIC 2007*. Also in: LNCS, 2007.

Ferrucci, D., Lally, A., UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*. 10(3-4) 2004.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., Lexical Markup Framework (LMF). In *Proceedings of LREC2006*, Genova, pp. 233-236. ELRA, Paris, 2006.

Gulrajani, G., Harrison, D., SHAWEL: Sharable and Interactive Web-Lexicons. In *Proceedings of the LREC2002 Workshop on Tools and Resources in Field Linguistics*, Las Palmas, pp. 1-4, 2002.

Huang, C.R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., Prévot, L. (eds.), *Ontologies and the Lexicon*. Cambridge Studies in NLP. Cambridge University Press, Cambridge, to appear.

Ide, N., Calzolari, N., Introduction to the Special Inaugural Issue. *Language Resources and Evaluation*. Springer, 39(1), pp. 1-7, 2005.

Ishida, T., Language Grid: An Infrastructure for Intercultural Collaboration. In *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100, 2006.

Kemps–Snijders, M., Nederhof, M., Wittenburg, P., LEXUS, a web-based tool for manipulating lexical resources. In *Proceedings of LREC2006*, Genova, pp. 1862-1865. ELRA, Paris, 2006.

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., Bertagna, F., Monachini, M., Soria, C., Calzolari, N., Huang, C.R., Hsieh, S.K., Towards an Architecture for the Global-WordNet Initiative. In *Proceedings of SWAP-06, 3rd Semantic Web Workshop*. 2006.

Monachini, M., Calzolari, N., Choukri, K., Friedrich, J., Maltese, G., Mammini, M., Odijk, J., Ulivieri, M., Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In *Proceedings of LREC2006*, Genova, pp. 1852-1857. ELRA, Paris, 2006.

Monachini, M., Quochi, V., Ruimy, N., Calzolari, N., Lexical Relations and Domain Knowledge: The BioLexicon Meets the Qualia Structure. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*. Paris, 2007.

Quochi, V., Del Gratta, R., Sassolini, E., Monachini, M., Calzolari, N., Toward a Standard Lexical Resource in the Bio Domain. In Vetulani, Z. (ed.), *Proceedings of 3rd Language and Technology Conference*, Poznań, pp. 295-299, 2007.

Roventini A., Ruimy N., Marinelli R., Ulivieri M., Mammini M., Mapping Concrete Entities from PA-ROLE-SIMPLE-CLIPS to ItalWordNet: Methodology and Results. In *Proceedings of the 45th Annual Meeting of the ACL*, pp. 161-164. Prague, 2007.

Soria, C., Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Huang, C., Calzolari, N., Towards agent-based cross-lingual interoperability of distributed lexical resources. In *Proceedings of COLING-ACL Workshop on Multilingual Lexical Resources and Interoperability*, Sydney, 2006.

Tokunaga, T., Sornlertlamvanich, V., Charoenporn, T., Calzolari, N., Monachini, M., Soria, C., Huang, C., Prevot, L., Xia, Y., Yu, H., Kiyoaki, S., Infrastructure for standardization of Asian language resources. In *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions*, pp. 827-834. Sydney, 2006.

Toral, A., Monachini, M., Formalising and bottom-up enriching the ontology of a Generative Lexicon. In *Proceedings of RANLP07 - Recent Advances in Natural Language Processing*. Borovets, Bulgaria, 2007.

Van Assem, M., Gangemi, A., Schreiber, G., Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of LREC2006*, Genova. ELRA, Paris, 2006.

A Flexible Framework for Integrating Annotations from Different Tools and Tagsets

Christian Chiarcos
Universität Potsdam
chiarcos@ling.uni-potsdam.de

Stefanie Dipper
Ruhr-Universität Bochum
dipper@linguistics.rub.de

Michael Götze
Universität Potsdam
goetze@ling.uni-potsdam.de

Julia Ritz
Universität Potsdam
julia@ling.uni-potsdam.de

Manfred Stede
Universität Potsdam
stede@ling.uni-potsdam.de

Abstract

Tools for linguistic annotation employ different data models and accompanying visualization metaphors, depending on the particular type of annotation envisaged. When a corpus is to be annotated on multiple layers, and the annotations are to be related to one another, the output formats of the annotation tools need to be unified. We describe an implemented framework for this step: reading the output of a variety of tools into a single database, where the data can be visualized, queried and evaluated across the layers. Then, besides the integration of resources at *format* level, we also seek compatibility between annotation *tagsets*: We describe how ontologies can be used to mediate between competing tagsets intended to cover the same class of linguistic phenomena.

1 Introduction

Manual linguistic annotation is labour-intensive and expensive. It is therefore of utmost importance to provide software environments that ensure the efficiency of the overall process. This can be done in three different ways: (1) by careful selection of the data to be annotated; (2) by providing (partial) automatic analyses that only need to be confirmed or changed by the human annotator; or (3) by tailoring the data models and the look-and-feel of annotation tools as good as possible to the kind of annotation performed.

In this paper, we are focusing on the last option. Nowadays, a variety of annotation tools are freely available, which support different styles of annotation for different purposes, such as layer-based transcription or labelling of words/phrases, coreference links, syntax trees, or discourse trees.

Another trend that has emerged in recent years is the availability of corpora annotated simultaneously on various levels, so that inter-relationships between the annotations can be explored. When such multi-layer corpora are to be created with existing dedicated annotation tools, a new problem arises: Output formats of the annotation tools can differ considerably, and annotations need to be aligned in order to be useful for purposes such as those mentioned above. To solve these problems, we have developed a software framework involving a generic standoff representation format; conversion from tool output to the generic format; aligning the annotations in a database that allows for visualization (which is not covered in this paper), retrieval, and statistical analyses of the data. By integrating an ontology in the query mechanism, resources based on differing annotation schemes can be queried simultaneously.

2 A Generic Standoff Format for Integrating Annotations

2.1 Representational standards

Nowadays the need for standardized annotation schemes and representation formats is widely recognized. Language resources must be well-documented and annotations be easy to interpret if they are to be beneficial for users other than the cor-

pus developers themselves. Standardization of representation formats concerns both the *physical* and *logical* data structures (see, e.g., (Schmidt, 2004)).

The logical data structure refers to the *data models* that are used to model the linguistic phenomena and their properties. We distinguish three types of data structures: (i) "annotation graphs": labeled directed acyclic graphs (LDAGs) whose nodes refer to a time line; annotation graphs are typically used for modeling time-aligned information (Bird and Liberman, 2001); (ii) structural annotations: DALGs whose nodes refer to other nodes; usually used for syntactic and other tree-like annotations; (iii) feature structures, used, e.g., for syntactic analyses in frameworks such as HPSG and LFG, but rarely used in the context of corpus annotation.

The division between the paradigms of time-aligned annotation graphs and hierarchical structures has weakened in recent years. For instance, the data model of annotation graphs has been generalized, resulting in the format ATLAS (Laprun et al., 2002), which supports both annotation graphs and hierarchical structures. Similarly, the NITE Object Model (Carletta et al., 2003b) and the general-purpose Linguistic Annotation Framework (Ide et al., 2003) serve both camps.

The physical data structure, on the other hand, refers to the "external" representation of the data. Here the de-facto standard is XML for serializing the data. Often, a standoff-architecture is used, which stores primary data and its annotations in different files (Sperberg and Bernard, 1994; Dybkjær et al., 1998). For the serialization of structural annotations, a natural way to represent trees is by using XML embedding structures. If structural annotations contain non-tree-like structures (e.g. crossing branches for discontinuous constituents), extra means like `xlink` attributes have to be employed (König and Lezius, 2000). Such representational means are less perspicuous and harder to interpret than the straightforward representation via XML embedding.

2.2 Integration of multiple annotations

Whereas these data models and formats might in principle host multi-level, heterogeneous annotation, projects that actually deal with data annotated at more than two levels (like MULI (Baumann et al., 2004)) tend to develop task-specific formats. Only recently, researchers started to integrate and merge annotations from different sources into one format: (Witt et al., 2005) merge multiple XML annotations of the same primary data into one XML format, leaving the original annotations intact as far as possible. For the representation of structurally-conflicting markup, elements have to be broken up and transformed into milestones. In contrast, (Ide and Suderman, 2007) propose one common pivot standard format, "GrAF", which all annotations have to be mapped onto. The format makes use of generic XML element names such as `node` and `edge` and encodes feature-value annotations by generic XML attributes `name` (e.g. "cat") and `value` (e.g. "NN").

In our approach, we pursue the same strategy as (Ide and Suderman, 2007). Our representation format, which we describe in the next section, is quite similar to the GrAF format. It serves as the "neutral" interchange format between different types of annotation structures and, at the same time, as the common import format to the linguistic database ANNIS (see Sct. 4). It supports querying and visualizing the data and its multi-level annotation, and includes ontology-based query evaluation which allows for searching data annotated with different tagsets. This integrated architecture probably distinguishes our approach from the above-mentioned ones.

2.3 Our representation format

Our representation format PAULA[1] (a German acronym for 'Potsdam interchange format for linguistic annotation') focuses on the integration of different annotation structures. We assume that corpus developers apply specialized annotation tools which are tailored to the specific annotation tasks. For instance, *annotate* (Brants and Plaehn, 2000) is frequently used for syntactic annotations; *Palinka* (Orasan, 2003) or *MMAX2*[2] for discourse-level annotations such as co-reference; *Exmaralda* (Schmidt, 2004) is applied for dialog transcription and various layer-based annotations. For these tools, we provide scripts that map the tool output to our representation format. The scripts are publicly available via the Internet: users can upload their data and

[1] http://www.sfb632.uni-potsdam.de/projects/d1/paula/doc/
[2] http://mmax2.sourceforge.net/

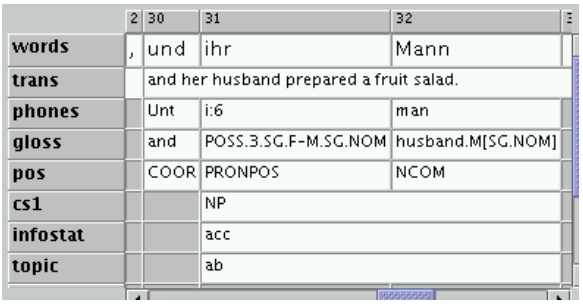


Figure 1: Annotation example (screenshot of the tool *Exmaralda*).

annotations (we currently provide converters for Exmaralda, MMAX2, Tiger XML (König and Lezius, 2000), URML (Reitter and Stede, 2003), Palinka, and a generic importer for annotations using inline-XML markup). The data is converted automatically to PAULA, and the user can copy it to the database ANNIS or perform statistical analyses with our WEKA-based application, see Sct. 4.3).

The mappings from the tool outputs to our format are defined such that they only transfer the annotations from one format into another without *interpreting* them or adding any kinds of information.

As an example, consider the original annotation of a short text fragment, annotated with the tool *Exmaralda*. Fig. 1 shows selected annotation levels, as displayed by the annotation tool[3]. Exmaralda's XML representation format implements annotation graphs, i.e., the primary data and all annotations refer to a common timeline, marked by timeline items (`tli`), whose IDs serve as anchors for the annotations. Annotations are called `event`s, they are anchored to the timeline via `start/end` attributes. The `tier` element specifies the type of annotation (e.g. `pos`), the `event` tags contain the actual annotation (`PRONPOS` for possessive pronoun, `NCOM` for common noun). The following fragment displays the primary data `ihr Mann` ('her husband') and their POS annotations.

```
<tli id="T18"/>
<tli id="T19"/>
<tli id="T44"/>
...
<tier id="TIE1" category="words">
<event start="T18" end="T19">ihr</event>
<event start="T19" end="T44">Mann</event>
<...
<tier id="TIE13" category="pos">
<event start="T18" end="T19">PRONPOS</event>
<event start="T19" end="T44">NCOM</event>
```

The corresponding representation of our pivot format PAULA presents the primary data in a `body` element. It defines `mark`ables for segments that receive annotations. A first layer of `markables` points to text regions in the `body` element, by means of `XPointer` expressions (see the markables with IDs `tok_20/21`). These markables can be thought of as tokens. Another layer of markables is added on top of the token markables (see the markables with IDs `pos_15/16`); they point to the tokens by means of `xlink:href` attributes. The actual annotations "PRONPOS" and "NCOM" are encoded by `feat` elements ("features"), which are anchored to the second layer of markables.

```
<body>... ihr Mann ...</body>
...
<mark id="tok_20" xlink:href="#xpointer(
   string-range(//body,'',97,3))"/>
   <!-- ihr -->
<mark id="tok_21" xlink:href="#xpointer(
   string-range(//body,'',101,4))"/>
   <!-- Mann -->
...
<mark id="pos_15" xlink:href="#tok_20"/>
<mark id="pos_16" xlink:href="#tok_21"/>
...
<feat xlink:href="#pos_15" value="PRONPOS"/>
<feat xlink:href="#pos_16" value="NCOM"/>
...
```

The reason for introducing an extra layer of `markable`s is that annotations in Exmaralda can refer to *spans* of token markables. In this case, there are two choices: Either anchor `feat` elements to a sequence of markables, similar to token markables, which are anchored to sequences of characters. Or introduce another layer of markables that are anchored to sequences of other markables, and `feat` elements then refer to this extra layer.

In principle our format could host both alternatives. However, we opted for the second alternative because we aim at rather rigid mapping "rules" so that the resulting pivot representation is as uniform as possible. This facilitates further processing and interpretation of the data.

[3]Layers (from top): the primary-data layer; translation to English; phonetic transcription according to Sampa; morpheme glosses, parts of speech, basic syntactic constituents ("cs1"), and information-structural annotation ("infostat", "topic") according to (Dipper et al., 2007).

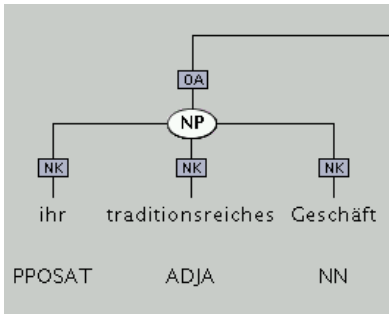


Figure 2: Annotation example (screenshot of the tool *TIGERSearch*).

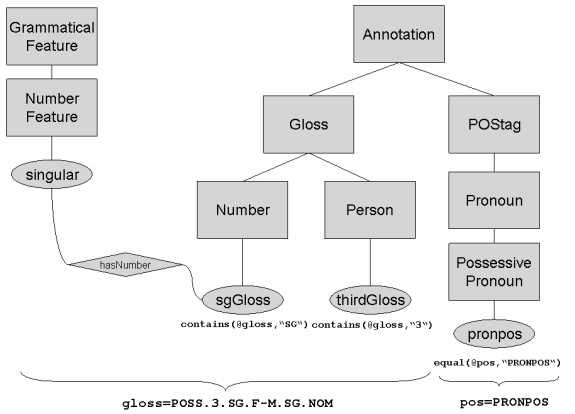


Figure 3: Fragment of Dipper et al.'s (2007) annotation model.

Another annotation example is shown in Fig. 2[4]. The annotations follow the STTS (Schiller et al., 1999) and TIGER (Brants et al., 2004) schemes.

3 An Ontology of Linguistic Annotations

So far, we have described aspects of the technical integration of multi-layered annotations from different sources and their representation. However, the integration of data from different sources (and partially from different languages) not only involves the integration of technical formats but also the conceptual integration. It is well-known that tag identifiers can differ widely and quite often involve idiosyncratic abbreviations. As an example, consider the great variety of tags assigned to *her* as a possessive determiner in different tag sets for English, which at a first glance seem to be fairly arbitrarily chosen at least in parts: `PP$` (Brown, (Greene and Rubin, 1981)), `TB` (London-Lund Corpus, (Eeg-Olofsson, 1991)), `PRP$` (Penn, (Santorini, 1990)), `DD` (POW, (Souter, 1989)), `PRON(poss,sing)` (ICE, (Greenbaum, 1992)), `APPGf` (Susanne, (Sampson, 1995)).

Here, we present a structured, modular ontology that is capable of both the conceptual integration of different annotation schemes by specifying a terminological reference, and the lossless representation of specific annotations.

This structured ontology involves two primary modules, a set of ANNOTATION MODELS which are representations of one annotation scheme, each, and a REFERENCE MODEL which represents a generalization over different annotation models, and thus, a common terminological reference.

A given annotation model is constructed solely on the basis of available annotation documentation, mostly guidelines if available, and annotated examples. Hence, it is a formalization of the annotation documentation, exhaustive with respect to the available documentation, but without any additional interpretation in terms of generally assumed linguistic categories, etc.

The partial ontological representation of the `pos` and `gloss` annotations of *ihr*, the German equivalent to the possessive pronoun 'her' (cf. Fig. 1) in terms of Dipper et al.'s (2007) model is given in Fig. 3. In the same way, annotations of the STTS tagset are represented in a separate annotation model.

While an annotation model is specific to one particular language, community, or purpose, the reference model is a general terminological resource, and consequently based on a broad range of resources, including specific annotation models, grammatical references, textbooks, but also existing terminological references such as the EAGLES recommendations for Morpho-Syntax (Leech and Wilson, 1996), and the GOLD ontology (Farrar and Langendoen, 2003). In case of divergent conceptualizations, e.g. the classification of attributive possessive pronouns

[4]The phrase *ihr traditionsreiches Geschäft* 'her traditional business' is annotated as an NP which functions as an accussative object ("OA"). Terminal nodes are labeled by POS tags according to the STTS tagset: "PPOSAT" (possessive pronoun), "ADJA" (attributive adjective), "NN" (common noun)

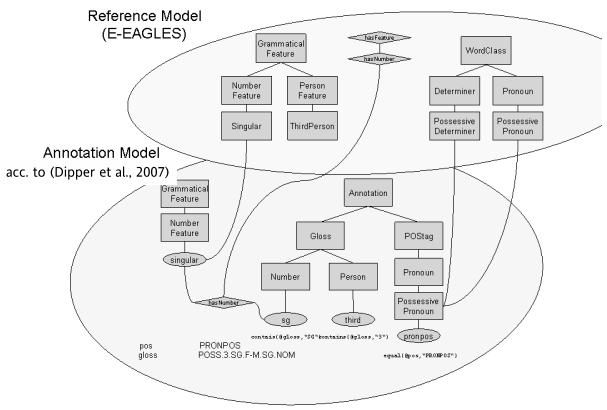


Figure 4: Fragment of E-EAGLES reference model and its linking with Dipper et al.'s (2007) annotation model.

as either Pronouns or Determiners, the EAGLES taxonomy was taken as an orientation, hence, the reference model is also referred to as *E(xtended)-EAGLES ontology*.

Annotation models and the reference model represent self-contained ontologies on their own. The conceptual integration of annotation models is then performed by means of a declarative LINKING between both the reference model and a specific annotation model. In the linking, every concept (class) of the annotation model is assigned a superclass from the reference model – including complex superclasses composed with the set operators $\cup, \cap$, or $\setminus$.

For the annotation model fragment in Fig. 3, the corresponding linking of concepts and the property `hasNumber` with their respective counterparts in the reference model is illustrated in Fig. 4.

In consequence of the linking, the concise annotation of *ihr* ('her', an example from Figures 1 and 2) can be rephrased in terms of the reference model. Consequently, an ontological description such as `PossessivePronoun and hasNumber(Singular) ...` naturally expands (by means of $\subseteq$ and $\in$) into a disjunction of several specific annotations according to different annotation models, e.g. matching both the scheme A tag `PRONPOS` and the scheme B tag `PPOSAT`.

The advantage of this structured account is that it avoids the plain identification of categories from different annotation schemes with standard categories, as it was required in older standardization approaches, e.g. (Leech and Wilson, 1996). Instead, the relations of high complexity can be specified and the necessary *interpretation* of categories in the annotation scheme is represented in an explicit, transparent, and modifiable way.

This tripartite structure of annotation models, reference model, and the linking in between can be augmented by the optional linking of the reference model with additional EXTERNAL REFERENCE MODELs, ontological formalizations of community- or language-specific terminological systems. Currently, we provide a linking with two external reference models, GOLD, the General Ontology of Linguistic Description (Farrar and Langendoen, 2003), developed in the context of language documentation, and the OntoTag ontologies (de Cea et al., 2004) developed in the context of Semantic Web applications, but so far specific to Romance languages.

We claim that this modular approach is more flexible as it allows alternative specifications of linking and the inclusion of alternative upper models as well as additional domain models. In contemporary annotation practice, its technological counterpart is the standoff paradigm (see Section 2).

4 A Database for Multiply Annotated Corpora

Having discussed both technical and conceptual issues of data integration, we now turn to the task of accessing integrated, multi-level corpora.

ANNIS[5] is a linguistic database that can be accessed as a server with standard web browsers via the internet, or installed on a local computer. The "local" ANNIS is a Java servlet application without a database backend: Operations for querying and visualizing are conducted on the data in main memory. This eases installation, but obviously limits the amount of data to be handled. The server version is currently being extended by a relational database. The overall goal of ANNIS is to provide access to heterogeneous multi-level annotations by providing suitable means both for visualization (which we do not address in this paper) and for querying.

[5] http://www.sfb632.uni-potsdam.de/annis/

Target user groups are linguists from different linguistic communities with basic computer skills, to whom developing or adapting existing query and visualization toolkits such as GATE (Cunningham et al., 2002) or the NITE XML Toolkit (Carletta et al., 2003a) would be too advanced or time-consuming. By providing import facilities for the PAULA pivot format described in Section 2, ANNIS supports the idea of distributed annotation with specialized ready-to-use tools. At present, our usage scenarios include the development and analysis of historical corpora, construction of a typological database with data from 16 different languages (Götze et al., 2005), and the creation of a text corpus with rich discourse-related annotations (Stede, 2004).

In the following sections, we focus on the facilities for querying and analyzing cross-layer phenomena that our system provides. The resources we chose for illustration in this paper are listed in Table 1. Corpus A is transcribed speech (maptask dialogues and question–answer pairs); corpus B is newspaper text, partially annotated by two annotators. The annotation layers given here are Information Structure (IS), Part-of-Speech (PoS) and Syntax; tools/formats are given in subscripts.

Corpus A	Corpus B
$\text{IS}_{EXMARaLDA}$	IS_{MMAX}
$\text{PoS, Syntax}_{EXMARaLDA}$	$\text{PoS, Syntax}_{TIGER}$

Table 1: Resources in our Database.

4.1 Annotation-based Querying

The query language implemented with ANNIS builds upon existing query languages and offers typical relations like dominance, inclusion ('_i_'), and overlap. Specifically, the language provides operators both for hierarchical and temporal relations. The latter are of particular relevance for querying multi-level annotations, since time often constitutes the only relation between annotations of different annotation levels. Moreover, the query language allows accessing different annotations of the same corpus, so that, for instance, competing analyses indicating disagreements between annotators can be found, as in (1) wrt. to the givenness of an item:[6]

(1) `ann1::givenness=new &`
`ann2::givenness=giv & #1 _=_ #2`

(2) `aboutness=ref & !givenness=* &`
`#1 _=_ #2`

The negation operator '!' allows us to formulate queries that check for the completeness of annotations. This is illustrated with (2), which checks (across layers) whether all referring expressions are annotated for the feature *givenness*.

4.2 Concept-based Querying

For cases where users are searching for instances of a certain annotation concept (see Section 3) rather than of a concrete tag, we provide for more abstract queries. A query preprocessor retrieves all tag descriptions that correspond to an ontological description and translates them into a disjunction of specific annotation values. If multiple annotation schemes (domain models) are considered, such a description may be expanded into a disjunction of tags from different tagsets and/or tiers.

Ontology-sensitive sub-queries are composed according to the following context-free grammar[7]:

ONTOQUERY := {CUE `in` ONTOEXP}
ONTOEXP := ONTOCONCEPT |
(ONTOEXP ONTOOP ONTOEXP) |
ONTOPROPERTY(ONTOFEATURE)
ONTOOP := `and` | `or` | `without`

Consequently, multiple queries for PoS tags from different annotation schemes can be replaced by one single ontology-sensitive corpus query. The query for possessive pronouns can be abbreviated as in (3).

(3) `pos in {PossessivePronoun}`

As opposed to the choice of regular expressions, this ontology-driven tag expansion allows a user to generalize over the specific form of annotations and tag names; it merely requires conceptual understanding.

[6] The queries (1) and (2) specify constraints over the annotations (givenness=new), their annotation set (ann1), and their relation ('_=_' states that both arguments refer to the same primary data). As (2) shows, wildcards can be used.

[7] ONTOCONCEPT, ONTOPROPERTY and ONTOFEATURE correspond to word classes, properties and grammatical features specified in the reference model. ONTOQUERYs can be embedded in arbitrary code which remains untouched during query expansion.

4.3 Cross-resource cross-layer analysis: Use cases

In addition to interactive queries, ANNIS provides for carrying out a range of statistical analyses.

Hypothesis testing. Suppose we want to investigate how givenness[8] of NPs is linked to their type. We enter a suite of concept-based queries similar to query (4) and receive a contingency table like Table 2.

(4) `givenness=giv & pos in {PossessivePronoun} & #1 _i_ #2`

	giv	acc	new
possNP	5	13	15
dem/defNP	133	91	135
indefNP	151	245	240
name	81	68	163
pers/demPron	109	22	8

Table 2: Contingency table: givenness vs. NP type

As a null hypothesis, we could assume stochastic independence between the features *givenness* and *NP type*. Using Pearson's (1900) χ^2, however, we can discard this hypothesis with high significance ($\chi^2 = 200.51, df = 8, p < .0005$).

Annotation Mining. We also exploit these merged resources to "re-feed" the annotation process by training classifiers on them. For this purpose, we built a component that exports data from the pivot format (see Sct. 2) to the Attribute Relation File Format (ARFF) used in WEKA (Witten and Frank, 2005), a common, ready-to-use data mining environment. As to the export, the user can specify the basic entity to be used, e.g. tokens, noun phrases, etc. Then, these entities are extracted (along with the features they are annotated with), forming one dataset per entity. In WEKA, experiments with different classifiers (SVMs, HMMs, decision trees) can be carried out. Currently, a reimport of the classification results to our pivot format is under construction. Thus, the automatic annotation can be presented to human annotators for correction.

[8](Dipper et al., 2007) use the values `giv`(en) for previously mentioned discourse referents, `acc`(essible) for referents that can be inferred from the context via 'bridging', `new` for referents new to recipient and discourse; non-referring NPs are not annotated.

5 Summary

We gave an overview of our software environment for producing multi-layer annotated corpora: a pivot format serving as "interlingua" between annotation tools, an ontology-based approach for mapping between tagsets, and a database that integrates the various annotations, and allows for querying the data (either by posing simple queries or by using the ontology) and for statistical analyses. Our conversion tools (to and from the pivot format) and the ANNIS database are freely available for research purposes. At present, we are adding a relational database to the server version of ANNIS. Future work will focus on improving our visualization of query results.

References

S. Baumann, C. Brinckmann, S. Hansen-Schirra, G. Kruijff, I. Kruijff-Korbayová, S. Neumann, and E. Teich. 2004. Multi-dimensional annotation of linguistic corpora for investigating information structure. In *Proc. of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, Boston.

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.

T. Brants and O. Plaehn. 2000. Interactive corpus annotation. In *Proceedings of LREC 2000*, Athens, Greece.

S. Brants, S. Dipper, P. Eisenberg, S. Hansen, E. König, W. Lezius, C. Rohrer, G. Smith, and H. Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. 2003a. The NITE XML Toolkit. *Behavior Research Methods, Instruments, and Computers*, 35(3).

J. Carletta, J. Kilgour, T. O'Donnell, S. Evert, and H. Voormann. 2003b. The NITE Object Model library for handling structured linguistic annotation on multimodal data sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web (NLPXML-2003)*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Meeting of the ACL*.

G. Aguado de Cea, A. Gómez-Pérez, I. Álvarez de Mon, and A. Pareja-Lora. 2004. Ontotag's linguistic ontologies: Improving semantic web annotations for a better language understanding in machines. In *ITCC '04: Proc. of the Int'l Conference on Information Technology: Coding and Computing*, Washington, DC, USA. IEEE Computer Society.

S. Dipper, M. Götze, and S. Skopeteas, editors. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*, volume 7 of *Interdisciplinary Studies on Information Structure (ISIS)*. Universitätsverlag Potsdam, Potsdam, Germany.

L. Dybkjær, N. Bernsen, H. Dybkjær, D. McKelvie, and A. Mengel. 1998. The MATE markup framework. MATE Deliverable D1.2.

M. Eeg-Olofsson. 1991. *Word-class tagging: Some computational tools*. Ph.D. thesis, Department of Linguistics and Phonetics, University of Lund, Sweden.

S. Farrar and D. T. Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100.

M. Götze, S. Skopeteas, T. Roloff, and R. Stoel. 2005. Towards a cross-linguistic production data archive: Structure and exploration. In Balder ten Cate and Henk Zeevat, editors, *TbiLLC*, volume 4363 of *Lecture Notes in Computer Science*, pages 127–138. Springer.

S. Greenbaum, 1992. *The ICE tagset manual*. University College London.

B. Greene and G. Rubin, 1981. *Automatic grammatical tagging of English*. Department of Linguistics, Brown University, Providence, R.I.

N. Ide and K. Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proc. of The Linguistic Annotation Workshop (LAW) 2007*, Prague.

N. Ide, L. Romary, and E. de la Clergerie. 2003. International Standard for a Linguistic Annotation Framework. In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology*.

E. König and W. Lezius. 2000. A description language for syntactically annotated corpora. In *Proc. of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 1056–1060, Saarbrücken.

C. Laprun, J. Fiscus, J. Garofolo, and S. Pajot. 2002. A practical introduction to ATLAS. In *Proceedings of LREC 2002*, Las Palmas,Spain.

G. Leech and A. Wilson, 1996. *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora*. Istituto di Linguistica Computazionale, Pisa.

C. Orasan. 2003. Palinka: a highly customisable tool for discourse annotation. In *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, Sapporo.

K. Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, 50(Series 5)*.

D. Reitter and M. Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In *Proc. of the 4th Int'l Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest, Hungary.

G. Sampson. 1995. *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Clarendon, Oxford.

B. Santorini, 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. Department of Computer and Information Science, University of Pennsylvania. Technical report MS-CIS-90-47.

A. Schiller, S. Teufel, C. Stöckert, and C. Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart and University of Tübingen.

T. Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*, Paris. ELRA.

C. Souter. 1989. A Short Handbook to the Polytechnic of Wales Corpus. Technical report, ICAME, Norwegian Computing Centre for the Humanities, Bergen University, Norway.

C. M. Sperberg and L. Bernard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Text Encoding Initiative, Chicago, Oxford.

M. Stede. 2004. The Potsdam Commentary Corpus. In *Proc. of the ACL Workshop on Discourse Annotation*, Barcelona.

A. Witt, D. Goecke, F. Sasaki, and H. Lüngen. 2005. Unification of XML Documents with Concurrent markup. *Literary and Linguistic Computing 2005*, 20:103–116.

I. Witten and E. Frank. 2005. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco, 2nd edition.

Supervised Clustering of the WordNet Verb Hierarchy for Systemic Functional Process Type Identification

Ian C Chow
Department of Chinese, Translation and Linguistics
City University of Hong Kong
ianchow@cityu.edu.hk

Jonathan J Webster
The Halliday Centre for Intelligent Applications of Language Studies
City University of Hong Kong
ctjjw@cityu.edu.hk

Abstract

This paper discusses a supervised lexicon-based semantic clustering for the purpose of constructing a verb lexicon based on Systemic Functional Linguistics (SFL). The work involves the interoperation of WordNet, FrameNet and SUMO. Selected WordNet verbs are annotated by SFL semantic category, i.e process type, by WordNet-FrameNet verbs mapping and FrameNet Frame-SFL Experience alignment. Our goal is to heuristically identify the process type of the un-annotated verbs by reference to annotated verbs. Automatic clustering is developed to locate the boundaries of semantic categories, and thereby avoid over-assimilation.

1 Introduction

Verb classification is an essential task for semantic analysis. NLP tasks involving event recognition, discourse polarity analysis and semantic role labeling require semantically categorized verbs lists for clause pattern and participant role identification. Verbs, serving as the nucleus of a clause, determine the event type and the semantic role of the entities involved in the event. Our aim is to construct a verb lexicon focusing on semantic categories. Moreover, the lexicon should be in machine readable form, in order to facilitate discourse analysis, semantic role labeling and meaning construal.

We employ the notion of Process Type, as employed in Systemic Functional Linguistics – SFL (Halliday, 2004; Halliday & Matthiessen, 1999), with a view to describing the construal of experience by means of natural language. An experience is typically realized grammatically in term s of a clause and depicts an event scenario. There are four main types of Process: Material, Mental, Verbal and Relational Processes, construing four types of experience. Each type of process possesses different semantic properties and lexicogrammatical features permitting different grammatical realizations in natural language. For example, in English, both a mental process and verbal process allow a that-clause as a participant argument, whereas a relational process allows only an adjective.

2 Interoperating Resources

We have made use of available resources including WordNet (Fellbaum, 1998), FrameNet (Fillmore et al, 2003), and Suggested Upper Merged Ontology – SUMO (Niles & Pease, 2001). WordNet provides intensive lexical coverage with semantic links among them but lacks information in clausal semantics. FrameNet identifies clause patterns, semantic role, verb argument structure and examples but a lower lexical coverage. SUMO is a non-linguistic upper ontology which has been mapped with WordNet (Niles & Pease, 2003). SUMO provides an encyclopedic concept hierarchy offering a non-linguistic semantic category for each WordNet synset.

WordNet is taken as the lexical resource. WordNet has a vast coverage of English words, 24890 verbs in 13650 synsets and are ontologically organized with various relation links. These links are useful tools for automatic semantic categorization. It is in machine readable format, mapping of various resources with WordNet are available and is currently in high attention and widely applied in the natural language processing.

2.1 FrameNet Frame mapping with SFL Experience

The issue of the commencement of SFL process identification is resolved by FrameNet-WordNet Mapping. The core concept of FrameNet, namely Frame, denotes an event scenario; each frame holds a list of lexical units (LU) comprised of words, mostly verbs, capable of evoking the frame. The event scenario denoted by a frame, in fact, conceptually describes instances of experience construal in SFL, thereby facilitating alignment between FrameNet Frames and SFL grammatical analysis. The alignment thus may be used to identify the process type of the verbs listed in LU; and by means of FrameNet-WordNet Mapping, WordNet verbs can be annotated according to their process type.

There are several works exploring mapping of FrameNet and WordNet. (Burchardt et al, 2005; Chow & Wong, 2006; Shi & Mihalcea, 2005) As our attention has focused on linking WordNet verbs to FrameNet LU, we have applied Shi & Mihalcea's (2005) mapping (hereafter, FnWnMap) which is verb focused. Verbs in FrameNet LU are tagged with WordNet sense. We further enhanced the mapping by assigning more WordNet verbs to the LU of the frames covered in the FnWnMap with a distributional statistical mapping utilizing the SUMO ontology (Chow & Webster, 2007)

Verbs from 314 frames were mapped in FnWnMap, 231 of them are mother classes and are taken out for alignment with SFL experience manually and the sub-class frames inherit the alignment from their mother-class. Special cases such as child frames with multi mother-classes do not exist in FnWnMap. However, there are 16 ambiguous frames which can be instances of more than one SFL type of experience. These 16 frames were neglected to avoid inheritance of the ambiguity on SFL experience to their LU verb list. 298 frames are thus each linked with a single SFL process type. Table 1 shows a portion of the alignment.

The interoperation of FrameNet, WordNet and SUMO has made it possible to annotate WordNet synsets by SFL process type. The result was then used as training data for further automatic identification of process type for the remaining WordNet data.

2.2 Extending process type identification

The main reason for the selection of WordNet as the primary lexical resource for our SFL verb lexicon is the ontological organization of WordNet data. The semantic links provided in WordNet offers a powerful tool for recruiting semantic related verb synsets as potential candidate for process type identification.

MATERIAL	RELATIONAL	VERBAL	MENTAL
FRAME	FRAME	FRAME	FRAME
absorb_heat abundance arraignment arriving assessing assistance atonement attention contrition corroding cooking_creation corroding_caused court_examination duplication departing escaping detaining differentiation dispersal event change_of_consistency committing_crime abusing kidnapping piracy rape robbery smuggling theft death	amounting_to bearing_arms coming_to_be compatibility dimension evaluative_comparison existence linguistic_meaning performers_and_roles position_on_a_scale possession containing similarity	attempt_suasion claim_ownership commitment communication statement gesture communication_noise communication_manner communication_response discussion evidence judgment_communication memory omen prevarication quarreling questioning reasoning reporting request sign speak_on_topic suasion talking_into telling topic	attempt seeking awareness certainty cogitation coming_to_believe deciding choosing desiring emotion_active expectation experiencer_subj experiencer_obj feeling hear judgment justifying perception appearance becoming_aware perception_experience predicting suspicion

Table 1. A portion of the mapping of FrameNet Frame and SFL Experience

Considering a process type annotated synset as a member of a prototype SFL process set, we attempt to recruit its related but process-unidentified synsets as candidate for the process prototype by verifying their semantic similarity with the annotated synset.

Taking process type as an attribute of the prototype set, all members in the class should belong to the same process type. In the other words, the process type will be assimilated to the recruited candidates possessing sufficient semantic similarity. The following section explains the recruitment of candidate synsets using WordNet links and the clustering of candidates for admission into the SFL process set.

3 Recruiting potential candidate for process identification

There are 11 links for verb synsets in WordNet: *Antonym, Hypernym, Troponym (as Hyponym in noun), Verb Group, Entailment, Cause, Also see, Derivationally related form, Domain of synset – TOPIC, Domain of synset – REGION, Domain of synset – USAGE.* Here we focus on the application of the Hypernym and Troponym links.

Hypernym and Troponym links make it possible to recruit synsets with high semantic relatedness. The links denote a sub-class relationship between synsets. In WordNet, the term troponym is used instead of hyponym, elaborating a more appropriate description for the is-a relation between verbs: *" X is one way to ?"* (hypernym) and, in reverse, *" X is a particular way to ?"* (troponym); for example, "build" has a hypernym "create"; "operate" has a troponym "drive". The links penetrate a number of synsets displaying a semantic prime along a continuum from general to specific, i.e. the is-a hierarchy of verbs. This distinguishes the two links by aggregating candidate in the form of A-B-C-D (A as an process annotated synset, B,C,D as the recruited candidates) from the other links demonstrating a connection of A-B (by *entailment* or *cause*) or A-B-A (as by *antonym*).

Disseminating process type among related synsets via hypernym-troponym link thus would increase the recall rate, and the semantic similarity provided merely by the link is preliminary reliable. For examples (troponyms are indicated by a "=>" sign):

Example (1) *[Mental Process]*:

{think, believe, consider, conceive} -- (judge or regard; look upon; judge; "I think he is very smart"; "I believe her to be very smart"; "I think that he is her boyfriend"; "The racist conceives such people to be inferior")

=>{rethink} -- (change one's mind; "He rethought his decision to take a vacation")

=>{feel} -- (have a feeling or perception about oneself in reaction to someone's behavior or attitude; "She felt small and insignificant"; "You make me feel naked"; "I made the students feel different about themselves")

=>{see, consider, reckon, view, regard} -- (deem to be; "She views this quite differently from me"; "I consider her to be shallow"; "I don't see the situation quite as negatively as you do")

=>{expect} -- (consider reasonable or due; "I'm expecting a full explanation as to why these files were destroyed")

=>{reconsider}-- (consider again; give new consideration to; usually with a view to changing; "Won't you reconsider your decision?")

Example (2) *[Material Process]*:

{touch} -- (make physical contact with, come in contact with; "Touch the stone for good luck"; "She never touched her husband")

=>{handle, palm} -- (touch, lift, or hold with the hands; "Don't handle the merchandise")

=>{manipulate} -- (hold something in one's hands and move it)

=>{operate, control} -- (handle and cause to function; "do not operate machinery after imbibing alcohol"; "control the lever")

=>{drive} -- (operate or control a vehicle; "drive a car or bus"; "Can you drive this four-wheel truck?")

4 Clustering the is-a hierarchy

The semantic relatedness offered by hypernym and troponym links allows aggregating a number of potential candidates. Nevertheless, assimilation of process type should not be disseminated along the whole is-a hierarchy arbitrarily, instead some restrictions are necessary. It is possible that there would be crossing of semantic category boundaries and thereby annotates an incorrect SFL process

type due to the continuum from generic to specific. Thus a supervised clustering along the is-a hierarchy is required in order to ensure the assimilation of process type is converted to a hypernym-troponym pair belonging to the same semantic category.

Along a is-a hierarchy, the more specific it is, the more likely a change of process type occurs. A typical example is the Material process verb synset {interact}, It has 20 troponyms and process type of which varies from Material process, Verbal process, Relational process, several troponyms are shown below indicated with a "=>" sign:

Example (3):

{Interact}*[Material Process]*-- (act together or towards others or with others; "He should interact more with his colleagues")

=>{communicate, intercommunicate} *[Verbal Process]* -- (transmit thoughts or feelings; "He communicated his anxieties to the psychiatrist")

=>{reach out} *[Verbal Process]*-- (attempt to communicate; "I try to reach out to my daughter but she doesn't want to have anything to do with me")

=>{communicate} *[Verbal Process]*-- (be in verbal contact; interchange information or ideas; "He and his sons haven't communicated for years"; "Do you communicate well with your advisor?")

=>{manipulate, keep in line, control} *[Material Process]* -- (control (others or oneself) or influence skillfully, usually to one's advantage; "She manipulates her boss"; "She is a very controlling mother and doesn't let her children grow up"; "The teacher knew how to keep the class in line"; "she keeps in line")

=>{have}*[Relational Process]*-- (have a personal or business relationship with someone; "have a postdoc"; "have an assistant"; "have a lover")

=>{invite, pay for} *[Material Process]* -- (have as a guest; "I invited them to a restaurant")

=>{socialize, socialize} *[Material Process]* -- (take part in social activities; interact with others; "He never socializes with his colleagues"; "The old man hates to socialize")

=>{get in touch, touch base, connect} *[Material Process]* -- (establish communication with someone; "did you finally connect with your long-lost cousin?")

=>{meet, gather, assemble, forgather, foregather} *[Material Process]* -- (collect in one place; "We assembled in the church basement"; "Let's gather in the dining room")

The process type of the synset {interact}, Material, should not be transmitted to all of its troponyms. To resolve this issue, we look for applicable information collected in our knowledge base – comprising WordNet, FrameNet and SUMO. Lexicographer file types from WordNet and SUMO concepts are applied to validate whether the troponym-hypernym pairs are in the same semantic category allowing process type transmission.

4.1 Lexicographer File Type

WordNet categorizes synsets into semantic categories known as lexicographer file types. There are 15 categories of verb synset. In our pilot investigation, some of the categories were found to be sub-categories of process type which can be taken as rules governing process identification; for example, all <*verb.contact*> would be identified as Material process. Most of the other file types are not able to be similarly applied; for example, <*verb.possession*> not only includes verbs of owning, which fall under the heading of Relational process, but also includes verbs of buying and selling, which are Material processes.

We take synset relation links as defining semantic relatedness between synsets. Adding the information about lexicographer file type, we are able to define the semantic distance between related synsets. If, for example, a hypernym-troponym pair belongs to the same type, their semantic relatedness is high, increasing their semantic similarity and vice versa.

In example (4) below, the dotted line represents the clustering boundary of synsets of different process type.

Example (4):

<verb.communication> {tell, evidence} -- (give evidence; "he was telling on all his former colleague")

=> <verb.communication> {inform}-- (act as an informer; "She had informed on her own parents for years")

=> <verb.communication> {inform} -- (impart knowledge of some fact, state or affairs, or event to; "I informed him of his rights")

=><verb.communication> {communicate, intercommunicate} -- (transmit thoughts or

feelings; "He communicated his anxieties to the psychiatrist")

clustering boundary

=><verb.social> {interact} -- (act together or towards others or with others; "He should interact more with his colleagues")

=> <verb.social> {act, move} -- (perform an action, or work out or perform (an action); "think before you act"; "We must move quickly"; "The governor should act on the new energy bill"; "The nanny acted quickly by grabbing the toddler and covering him with a wet towel")

A preliminary clustering is thus attained by the change of lexicography file type along the is-a hierarchy of verb synsets. However, further investigation found that merely using the lexicography file type leaves behind a number of appropriate candidate synsets. Example (5) shows a set of troponyms of the verb synset {change}, the un-recalled troponyms due to lexicography file type are indicated with an asterisk.

Example (5):

<verb.change> *[Material Process]* {change} -- (undergo a change; become different in essence; losing one's or its original nature; "She changed completely as she grew older"; "The weather changed last night")

=><verb.change> *[Material Process]* {form} -- (assume a form or shape; "the water formed little beads")

=><verb.change> *[Material Process]* {adjust, conform, adapt} -- (adapt or conform oneself to new or different conditions; "We must adjust to the bad economic situation")

*=><verb.cognition> *[Material Process]* {synthesize} -- (combine and form a synthesis)

*=>{01069806}<verb.competition> *[Material Process]* {promote} -- (be changed for a superior chess or checker piece)

=><verb.motion>[Material Process]* {settle} -- (become clear by the sinking of particles; "the liquid gradually settled")

*=><verb.perception> *[Material Process]* {solarize, solarise} -- (become overexposed; "The film solarized")

*=><verb.social> *[Material Process]* {liberalize, liberalise} -- (become more liberal; "The laws liberalized after Prohibition")

*=> <verb.social> *[Material Process]* {stratify} -- (develop different social levels, classes, or castes; "Society stratifies when the income gap widens")

*=><verb.social> *[Material Process]* {relax, loosen} -- (become less severe or strict; "The rules relaxed after the new director arrived")

*=><verb.stative> *[Material Process]* {stagnate} -- (cause to stagnate; "There are marshes that stagnate the waters")

4.2 SUMO Concept

A complementary approach is facilitated by the SUMO-WordNet Mapping. Each WordNet synset is mapped to a SUMO concept, denoting the semantic category of the concept definition of the synset. The semantic category provided by SUMO is motivated by, although non-linguistically, a world-knowledge encyclopedic concept taxonomy. We combine both the SUMO concept and the WordNet lexicography file type to determine the clustering boundary within the verb is-a hierarchy. A hypernym-troponym pair is determined to be clustered as the same process-type sharing category if they are not in the same lexicography file type but mapped to the same [SUMO concept].

Example (6):

[INJURING] <verb.body> {injure, wound} -- (cause injuries or bodily harm to)

=> [INJURING] <verb.contact> {bruise, contuse} -- (injure the underlying soft tissue of bone of; "I bruised my knee")

=> [INJURING] <verb.contact> {jam, crush} -- (crush or bruise; "jam a toe")

The concept hierarchy of SUMO is merged from several ontologies for the purpose of constructing a uniform and applicable upper ontology. It has a very delicate structure including 20,000 concept terms and 60,000 axioms. Comparing this with the number of concepts included in a general linguistic verb classification, the occurrences of equivalent SUMO concepts in a hypernym-troponym pair is less frequent than using lexicographer file type.

Example (7):
[*=> : un-clustered synset]
[%=>: clustered by <lexicographer file type>]
[#=>: clustered by [SUMO]]

[SEPARATING] <verb.contact> {separate} -- (divide into components or constituents; "Separate the wheat from the chaff")

*=> [DIALYSIS] <verb.change> {dialyse, dialyze} -- (separate by dialysis)

#=> [SEPARATING] <verb.change> {decompose, break up, break down} -- (separate (substances) into constituent elements or parts)

%=> [SEPARATING] <verb.change> digest -- (soften or disintegrate by means of chemical action, heat, or moisture)

%=> [SEPARATING] <verb.change> dissociate -- (to undergo a reversible or temporary breakdown of a molecule into simpler molecules or atoms; "acids dissociate to give hydrogen ions")

%=> [SEPARATING] <verb.change> crack -- (reduce (petroleum) to a simpler compound by cracking)

#=> [SEPARATING] <verb.change> peptize, peptise -- (disperse in a medium into a colloidal state)

#=> [SEPARATING] <verb.change> macerate -- (separate into constituents by soaking)

%=> [SEPARATING] <verb.contact> sift, sieve, strain -- (separate by passing through a sieve or other straining device to separate out coarser elements; "sift the flour")

%=> [SEPARATING] <verb.contact> rice -- (sieve so that it becomes the consistency of rice; "rice the potatoes")

%=> [SEPARATING] <verb.contact> resift -- (sift anew)

%=> [SEPARATING] <verb.contact> riddle1, screen1 -- (separate with a riddle, as grain from chaff)

%=> [SEPARATING] <verb.contact> winnow, fan -- (separate the chaff from by using air currents; "She stood there winnowing chaff all day in the field")

%=> [SEPARATING] <verb.contact> wash -- (separate dirt or gravel from (precious minerals))

%=> [SEPARATING] <verb.contact> pan, pan out, pan off -- (wash dirt in a pan to separate out the precious minerals)

#=> [SEPARATING] <verb.motion> avulse -- (separate by avulsion)

In the above example, all semantic clustered synsets (preceded by '%' or '#') will be identified by process type. In the actual processing of the data, for each verb, which is process annotated by FrameNet-WordNet mapping and Frame-Experience alignment, can generate a result of a whole clustered set of verbs being annotated by a SFL process type.

5 Conclustion

The construction of a SFL semantic verb lexicon is facilitated by semantic similarity determined heuristically by reference to WordNet, FrameNet and SUMO. In addition to mapping between these three resources, we have also incorporated information from SFG into FrameNet, specifically in terms of relating the SFG grammatical analysis of Experience with Frame Semantics.

This work particularly focuses on the power of semantic relatedness offered by WordNet links to locate a group of candidate verb synsets. Semantic relatedness itself is a particular kind of semantic similarity (Turney, 2006), such that producing a preliminary cluster, a prototype of process-type sharing set. In order to fine-tuning the prototype, information from WordNet and SUMO have been extracted and reused in order to eliminate clustered members with insufficient semantic similarity. Both linguistic and non-linguistic conceptualizations of knowledge were applied. To the linguistically motivated WordNet lexicographer file type, was added an encyclopedic conceptualization, SUMO. By doing so we were able to cluster WordNet is-a hierarchical semantic primes into sets along with verb synset members of sufficient semantic similarity.

The clustering approach is not only useful for semantic properties identification, as in our reported work, but also it can be a tool for measuring semantic similarity between concepts.

Semantic similarity measurement has been an interesting topic with high attention in natural language processing (Pederson et al, 2004; Banerjee & Pederson, 2003; Turney, 2006) Algorithm are mostly lexicon-based, corpus-based or hybrid of the two (Lesk, 1986, 1969; Bundanitsky and Hirst, 2001; Turney, 2005; Resnik 1995). This paper proposes an ontology mapping approach which relies on interoperable semantic resources.

References

Bateman, John A. 1991. “The Theoretical Status of Ontologies in Natural Language Processing” in the Proceedings of *the workshop on ‘Text Representation and Domain Modelling – Ideas from Linguistics and AI’. Berlin.*

Bateman, John A, Renate Henschel, Fabio Rinaldi. 2005. Generalized Upper Model 2.0, On-line publication at I1-OntoSpace: *http://www.ontospace.uni-bremen.de/twiki/bin/view/Main/LinguisticOntology*

A. Burchardt, K. Erk, and A. Frank. 2005. A WordNetdetour to FrameNet. *In Proceedings of the GLDV 2005 Workshop GermaNet II, Bonn.*

Banerjee, S., Pedersen, T. 2003. Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence pp. 805–810*

Bundanitsky, Alexander and Graeme Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. *In Proceedings of Worksjop on WordNet and Other Lexical Resources, NACCL-2001 pp29-34, Pittsburgh, PA.*

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. “Background to FrameNet” *International Journal of Lexicography 16: 235-250.*

Chow, Ian C. & Webster, Jonathan J. 2007. "Interfacing WordNet, FrameNet and SUMO". *In Proceedings of the GLDV 2007 Workshop Lexical-Semantics and Ontological Resources, Tübingen.*

Chow, Ian C. & Webster, Jonathan J. 2007. Integration of Lingtuistic Resource for Verb Classification: FrameNet Frame, WordNet Verb and Suggested Upper Merged Ontology. *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science Vol. 4394, pp1-11*

Chow, Ian C. & Webster, Jonathan J. 2006. "Mapping FrameNet and SUMO with WordNet Verb: Statistical Distribution of Lexical-Ontological Realization," *Proceedings of the Fifth Mexican International Conference on Artificial Intelligence (MICAI'06). pp. 262-268.*

Chow, Ian C. & Wong, Tak Ming. 2006. “Axiomatizing Relational Network for Knowledge Engineering - Exploring WordNet and FrameNet”. *In Proceedings of The 2006 IEEE International Conference on Information Reuse and Integration. Hawaii, USA. pp. 262-267.*

Fellbaum, Christiane. 1998. *WordNet An Electronic Lexical Database*. MIT Press. Cambridge. (1998)

Halliday, MAK; revised by Christian M.I.M Matthiessen. 2004. *An introduction to functional grammar*. London: Arnold

Halliday, MAK & Matthiessen, Christian MIM. 1999. *Construing experience through meaning*. New York: Continuum.

Ide, Nancy. 2006. Making Senses: Bootstrapping Sense-tagged Lists of Semantically-Related Words. In Gelbukh, Alexander (Ed.), *Computational Linguistics and Intelligent Text, Lecture Notes in Computer Science, Springer.*

Lesk, Michael E. 1969. Word-word associations in document retrieval system. *American Documentation, 20(1):27-38.*

Lesk, Michael E. 1986. Automatic sense dimambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *In Proceedigns of ACM SIGDOC’86, pp24-26, New York*

Levin, Beth. 1993. *English Verb Class and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

Maedche, Alexander D. 2002. *Ontology learning for the semantic Web*. Kluwer Academic Publishers. London.

Merlo and Stevenson. 2001. “Automatic Verb Classification Based on Statistical Distribution of Argument Structure”, *Computational Linguistics, 27:3, pp. 373--408.*

Niles, I., and Pease, A. 2001. Towards a Standard Upper Ontology. *In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17-19.*

Niles, I., and Pease, A. 2003. “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology”, *Proceedings of the IEEE International Conference on Information and Knowledge Engineering. pp. 412-416.*

Oltramari, A. 2006. “LexiPass methodology: a conceptual path from frames to senses and back” *Proceedings of the Fifth international conference on Language Resources and Evaluation, LREC.*

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intellifence (IJCAI-95), pp448-453, San Francisco, CA.*

Ruppenhofer, J., Ellsworth, M., Petruck, M.,, Johnson, C. 2006. FrameNet II: Theory and Practice. On-line publication at *http://framenet.icsi.berkeley.edu/*

Pedersen, T., Patwardhan, S., Michelizzi, J. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. *Proceedings of the Nineteenth National Conference on Artificial Intelligence pp.1024-1025*

Shi, L., Mihalcea, R. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and Word-Net for Robust Semantic Parsing. *Computational Linguistics and Intelligent Text Processing, 6th International Conference Proceedings. Lecture Notes in Computer Science, Vol. 3406. pp. 100-111*

SUMO. The Suggested Upper Merged Ontology. *http://www.ontologyportal.org/*

Turney, Peter D and Michael L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning, 60(1-3):251-278.*

Turney, Peter D. 2006. Similarity of Semantic Relations. *Computatiol Linguistics* 32(3): 379-416.

Creating an Interoperable Language Resource for Interoperable Linguistic Studies

Alex Chengyu Fang
Department of Chinese, Translation and Linguistics
City University of Hong Kong
acfang@cityu.edu.hk

Abstract

The International Corpus of English (ICE) is a corpus for the study of English as a global language. The project is parameterised by component, regional sub-corpora and a set of predefined textual categories. The one-million-word British component has been constructed, grammatically tagged, and syntactically parsed. This article is first of all a description of steps taken to ensure conformity within the project. These include corpus design, part-of-speech tagging, and syntactic parsing. The article will then present a study that examines the use of adverbial clauses across speech and writing, illustrating the necessity for interoperable analysis of linguistic data.

1 Introduction

The International Corpus of English (ICE) is a project that aims at the construction of interoperable language resources that enable systematic investigations of the grammatical properties of English as a global language. In particular, it aims at the construction of a collection of corpora for countries and regions where English is used either as a majority first language (such as Australia, Britain and the US) or as an additional official language (such as India and Singapore; Greenbaum 1992). Each variety should be represented through 500 samples of 2,000 words each, both spoken and written by adults of 18 and above who have received formal education through the medium of English to the completion of secondary school. The project comprises over 20 national or regioinal teams, each dealing with a component corpus. For such an ambitious project, it was important to have a set of clearly defined criteria to guide through the various stages of the project regarding corpus design, corpus annotation, and corpus analysis in order to ensure conformity and hence interoperability. According to Greenbaum (1996:5), "for valid comparative studies the components of ICE need to follow the same design, to date from the same period, and to be processed and analysed in similar ways".

The purpose of this article is two folded. It will first of all introduce the ICE project through issues related to corpus design and corpus annotation at grammatical and syntactic levels. It will then present a contrastive study of the use of adverbial clauses across speech and writing based on the British component of ICE. As will be shown, the results are contrary to previous findings. Explanations will be offered in the light of interoperable analysis issues.

2 Interoperable Corpus Design and Annotation

This section deals with the design of the corpus and the annotations applied to it. In particular, it will first describe the text composition of the corpus and then discuss POS tagging and syntactic parsing.

2.1 Corpus Design

The interoperability of the project was first handled at the level of corpus design. Each national or regional corpus was to be constructed according to an identical composition of text categories. Such a design is illustrated by Table 1, which summarises the types of texts to be represented in the project. As can be see, the corpus contains both spoken and written material. The spoken section comprises dialogues and monologues. The former is represented by a register that changes from the less formal setting such as direct conversions to a more formal setting such as legal cross-examinations and business transactions. The latter ranges from unscripted speech to mixed and scripted speech. The written section can also be described according to a continuum that starts from non-printed material comprising student essays and social letters to published and hence necessarily polished language used in both learned and popular writings.

Spoken				Written			
Dialogue	Private			Non-Printed	Student Writing		
	S1A1	direct conversations	90		W1A1	untimed essays	10
	S1A2	distanced conversations	10		W1A2	timed essays	10
	Public				Correspondence		
	S1B1	class lessons	20		W1B1	social letters	15
	S1B2	broadcast discussions	20		W1B2	business letters	15
	S1B3	broadcast interviews	10	Printed	Informational		
	S1B4	parliamentary debates	10		W2A1	Learned: humanities	10
	S1B5	legal cross-examinations	10		W2A2	Learned: social sciences	10
	S1B6	business transactions	10		W2A3	Learned: natural sciences	10
Monologue	Unscripted				W2A4	Learned: technology	10
	S2A1	spontaneous commentaries	20		W2B1	Popular: humanities	10
	S2A2	unscripted speeches	30		W2B2	Popular: social sciences	10
	S2A3	Demonstrations	10		W2B3	Popular: natural sciences	10
	S2A4	legal presentations	10		W2B4	Popular: technology	10
	Mixed				W2C1	Press news reports	20
	S2B1	broadcast news	20		Instructional		
	Scripted				W2D1	Administrative writing	10
	S2B2	broadcast talks	20		W2D2	Skills and hobbies	10
	S2B3	non-broadcast talks	10		Persuasive		
					W2E1	Press editorials	10
					Creative		
					W2F1	Fiction	20

Table 1: The composition of the ICE corpus

2.2 The ICE wordclass annotation scheme

The second measure taken to ensure interoperability within ICE for its subsequent analysis was the design of a standard scheme for word-class analysis. This standard is maintained through both a written manual (Greenbaum and Ni 1996) and an automatic part-of-speech (POS) tagging system that automatically applies such a standard to electronic texts (Fang 1996a). There are altogether 22 head tags and 71 features in the ICE wordclass tagging scheme, resulting in about 270 grammatically possible combinations. They cover all the major English word classes and provide morphological, grammatical, and collocational information. A typical ICE tag has two components: the head tag and its features that bring out the grammatical features of the associated word. For instance, **N(com,sing)** indicates that the lexical item associated with this tag is a common (**com**) singular (**sing**) noun (**N**).

Tags that indicate phrasal collocations include **PREP(phras)** and **ADV(phras)**, prepositions (as in [1]) and adverbs (as in [2]) that are frequently used in collocation with certain verbs and adjectives:

[1] *Thus the dogs' behaviour had been changed because they associated the bell with the food.*

[2] *I had been filming The Paras at the time, and Brian had had to come down to Wales with the records.*

Some tags, such as **PROFM(so,cl)** (pronominal *so* representing a clause as in [3]) and **PRTCL(with)** (particle *with* as in [4]), indicate the presence of a clause; *so* in [3] signals an abbreviated clause while *with* in [4] a non-finite clause:

[3] *If so, I'll come and meet you at the station.*

[4] *The number by the arrows represents the order of the pathway causing emotion, with the cortex lastly having the emotion.*

Examples [5]-[7] illustrate tags that note special sentence structures. *There* in [5] is tagged as **EXTHERE,** existential *there* that indicates a marked sentence order. [6] is an example of the cleft sentence (which explicitly marks the focus), where *it* is tagged as **CLEFTIT**. [7] exemplifies anticipatory *it*, which is tagged as **ANTIT**:

[5] *There were two reasons for the secrecy.*

[6] *It is from this point onwards that Roman Britain ceases to exist and the history of sub-Roman Britain begins.*

[7] *Before trying to answer the question it is worthwhile highlighting briefly some of the differences between current historians.*

The verb class is divided into auxiliaries and lexical verbs. The auxiliary class notes modals, perfect auxiliaries, passive auxiliaries, semi-auxiliaries, and semip-auxiliaries (those followed by *-ing* verbs). The lexical verbs are further annotated according to their complementation types. There are altogether seven types: complex-transitive, complex-ditransitive, copular, dimonotransitive, ditransitive, intransitive, monotransitive, and **TRANS**. Figure 1 shows the sub-categorisations of the verb class.

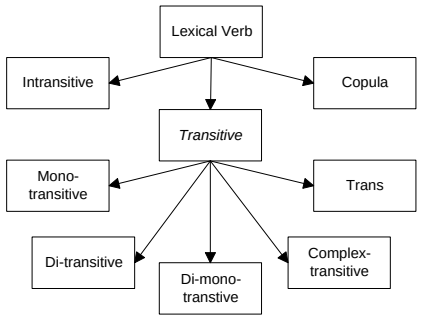


Figure 1: The ICE subcategorisation for verbs

The notation **TRANS** of the transitive verb class is used in the ICE project to tag those transitive verbs followed by a noun phrase that may be the subject of the following non-finite clause. This type of verb can be analysed differently according to various tests into, for instance, mono-transitives, ditransitives and complex transitives. To avoid arbitrary decisions, the complementing non-finite clause is assigned a catch-all term 'transitive complement' in parsing, and its preceding verb is accordingly tagged as **TRANS** in order to avoid making a decision on its transitivity type. This verb type is best demonstrated by [8]-[11]:

[8] *Just before Christmas, the producer of Going Places, Irene Mallis, had asked me to make a documentary on 'warm-up men'.*

[9] *They make others feel guilty and isolate them.*

[10] *I can buy batteries for the tape - but I can see myself spending a fortune!*

[11] *The person who booked me in had his eyebrows shaved and replaced by straight black painted lines and he had earrings, not only in his ears but through his nose and lip!*

In examples [8]-[11], *asked, make, see,* and *had* are all complemented by non-finite clauses with overt subjects, the main verbs of these non-finite clauses being infinitive, present participle and past participle.

As illustrated by examples [1]-[11], the ICE tagging scheme has indeed gone beyond the wordclass to provide some syntactic information and has thus proved itself to be an expressive and powerful means of pre-processing for subsequent parsing.

The ICE tagging scheme is automatically applied by AUTASYS, a part-of-speech tagging system that is fast (over one million words per minute), accurate (over 96% accuracy) and robust. See Fang (1996a) for more detailed descriptions,

2.3 The ICE parsing scheme

The third step taken to ensure interoperability is the design of a parsing scheme that handles the analysis of the corpus at syntactic level. This step is reinforced through an automatic system that applies the annotation scheme to texts that have already been POS tagged (Fang 1996b and 2000). The automatically produced sentences as parsed trees were then subject to manual manipulation through a graphical tree editor that maximally helps the editing of the trees through linguistically licensed constraints on category-function combinations. The ICE parsing scheme recognises five basic syntactic categories. They are adjective phrase (**AJP**), adverb phrase (**AVP**), noun phrase (**NP**), prepositional phrase (**PP**), and verb phrase (**VP**). Each tree in the ICE parsing scheme is re-presented as a functionally labelled hierarchy, with features describing the characteristics of each constituent, which is represented as a pair of function-category labels. In the case of a terminal node, the function-category descriptive labels are appended by the lexical item itself in curly brackets. Figure 2 is such a structure for [12].

[12] *We will be introducing new exam systems for both schools and universities.*

According to Figure 2, we know that [12] is a parsing unit (**PU**) realised by a clause (**CL**), which governs three daughter nodes: **SU NP** (NP as subject), **VB VP** (VP as verbal), and **OD NP** (NP as direct object). Each of the three daughter nodes are sub-branched until the leaves nodes with the input tokens in curly brackets. The direct object node, for example, has three immediate constituents: **NPPR AJP** (AJP as NP pre-modifier), **NPHD N(com,plu)** (plural common noun as the NP head), and **NPPO PP** (PP as NP post-modifier).

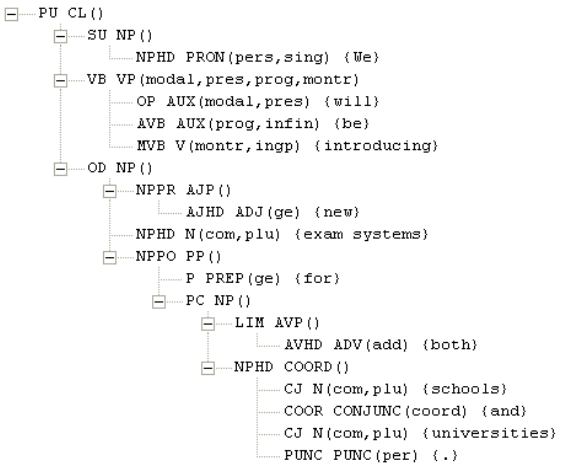


Figure 2: A parse tree for [12]

Note that in the same example, the head of the complementing NP of the prepositional phrase is initially analysed as a coordinated construct (**COORD**), with two plural nouns as the conjoins (**CJ**) and a coordinating conjunction as coordinator (**COOR**).

The ICE parsing scheme can be automatically applied to large quantities of natural text by the Survey Parser, a syntactic parsing system that is fast (over 50,000 words per minute), accurate (over 80% accuracy) and robust, producing either a full parse or a partial analysis without crashing. See Fang (1996b, 2000, and 2008) for more detailed descriptions.

3 Interoperable Linguistic Studies

The ICE corpus is thus an interoperable language resource that will maximally enable interoperable linguistic studies in, for instance, the grammatical differences and similarities of varieties of English. Since this resource is annotated according to grammatical and syntactic schemes that are theory neutral, it will also allow for comparisons with other resources.

In what follows, a study will be presented that aims at the investigation of the use of adverbial clauses across speech and writing based on the British component of ICE (ICE-GB). Results of the investigation will be presented and discussed in the light of findings of past studies. The experiments examined the frequency distribution of finite adverbial clauses as well as the non-finite ones (infinitival, present participial, and past participial) in ICE-GB. There are three procedures. First, the experiment aimed to establish the overall distribution of adverbial clauses across the spoken and the written sections. Secondly, samples of spontaneous and prepared speech were examined to ascertain whether preparedness could be seen as a continuum of changes for the use of adverbial clauses. Finally, samples of timed and untimed university essays were used to validate the hypothesis that adverbial clauses also demonstrate a predictable variation as a function of degrees of preparedness in written English.

3.1 Uses of adverbial clauses across speech and writing

As a first step, the complete corpus was used to obtain empirical indications of the different uses of adverbial clauses across speech and writing. Frequencies of occurrence were respectively collected from the spoken and the written sections of ICE-GB. The statistics include the total number of sentences and clauses in these two sections. Statistics were also collected for the total number of sentences involving the use of adverbial clauses and the exact number of adverbial clauses in these two sections. Two proportions were calculated: the total number of sentences with at least one adverbial clause over the total number of sentences, and the total number of adverbial clauses over the total number of sentences. The former indicates the proportion of sentences in ICE-GB that make use of adverbial clauses. The latter shows the proportion of adverbial clauses in the corpus since there often are multiple adverbial clauses in one sentence or utterance and it is useful to have such an indication. These two proportions thus indicate how often adverbial clauses are used and how complex the sentence structure is (assuming that structural complexity can be measured in terms of clause subordination). Table 2 summarises the results.

Table 2. Adverbial clauses in speech and writing

	Spoken (59,470)		Written (24,084)		Total (83,554)	
	#	%	#	%	#	%
Sentence	7124	11.98	6474	26.88	13598	13.27
Clause	7809	13.13	7052	29.28	14861	17.79

Initial results indicate that the uses of adverbial clauses are more frequent in writing than in speech. As Table 2 clearly indicates, a much higher proportion of sentences in writing make use of adverbial clauses. To be exact, adverbial clauses are more than twice likely to occur in writing than in speech. In writing, 25.42% of the sentences make use of adverbial clauses in contrast to only 12.49% of the sentences with an adverbial clause in speech. The same difference can be observed in terms of the number of adverbial clauses: there are over 30 adverbial clauses per one hundred sentences in

writing compared with fewer than 15 adverbial clauses per one hundred sentences in speech. Note that the proportions are normalised according to the number of sentences and clauses. It makes more sense in terms of sentences rather than words but even in terms of words speech demonstrates a smaller proportion of adverbial clauses than writing. As a general guide, there are 600,000 words in the spoken section of the corpus and 400,000 words in the written section. In terms of words, therefore, there are 1.46 adverbial clauses per hundred words in speech, compared with 1.86 in writing.

3.2 Types of adverbial clauses across speech and writing

The distribution of different types of adverbial clauses was investigated in order to verify that the observed difference was not the result of a skewed use of any one particular type. The second experiment examined the distribution of finite adverbial clauses with an overt subordinator and the non-finite ones, which include infinitival, present participial and past participial adverbial clauses. They are illustrated respectively by examples (2)-(5) with the relevant sections underlined.

(2) *And I think the question is bigger than that because it's from both sides.* <#S1A-001-054>

(3) *Having said that, I can really only say how it was for me when I came to work.* <#S1A-001-056>

(4) *And you condemn the series having seen a bit of one of them.* <#S1A-006-105>

(5) *The actual work surface was a very thick piece of wood, dumped on top, all held in place by words.* <#S1A-009-200>

The results are summarized in Table 3. As can be clearly seen, this second experiment also indicates that written samples of the ICE corpus make much more extensive use of the adverbial clause, be it finite, infinitival, or participial. The finite ones occur twice as many times in writing than in speech. For the other three types of adverbial clauses, the proportion for the written genre is even higher than for the spoken genre. Consider the infinitival clauses, for example. In writing, they are nearly three times more likely to be used than in spoken discourse (5.43% vs 1.98%), largely echoing previous observations that writing is characterised by a higher content of infinitives compared with spoken English (see, for example, [6] and [8]). This proportion is even greater with the other two types of non-finite adverbial clauses.

Table 3. Types of adverbial clauses across speech and writing

		Spoken (59,470)		Written (24,084)		Total (83,554)	
		#	%	#	%	#	%
A_{sub}	Sentence	5172	8.69	3954	16.42	9126	10.92
	Clause	5787	9.73	4430	18.39	10217	12.23
A_{infin}	Sentence	1122	1.89	1254	5.21	2376	2.84
	Clause	1177	1.98	1308	5.43	2485	2.97
A_{ing}	Sentence	691	1.16	1023	4.25	1714	2.05
	Clause	704	1.18	1066	4.43	1770	2.12
A_{edp}	Sentence	139	0.23	243	1.01	382	0.46
	Clause	141	0.24	248	1.03	389	0.47
Total	Sentence	7124	11.98	6474	26.88	13598	16.27
	Clause	7809	13.13	7052	29.28	14861	17.79

We may incidentally note that past participial clauses are the least frequent type of adverbial clauses, with only 141 found in speech and 248 in writing in the whole corpus.

3.3 Types of adverbial clauses across spontaneous and prepared speech

Empirical indications thus suggest that adverbial clauses are a marked characteristic of the written genre, in line with non-finite clauses that also characterise writing. However, to conclude that this difference in terms of use is due to different levels of elaboration, we need further empirical evidence. We need to demonstrate that such variations can be observed not only across speech and writing, but also within the spoken and the written sections as a function of varying degrees of elaboration.

To this end, a sub-corpus of 180,000 words was created with S1A texts in ICE-GB, representing spontaneous private conversations. A second sub-corpus was also created, this time with the first 40 texts in S2B, representing talks prepared and scripted for public broadcast. These two genres thus may be seen as forming a continuum between what was unprepared and what was carefully prepared, therefore a measure of different degrees of elaboration.

The results are summarised in Table 4, where we can read that, as an example, the subcorpus of spontaneous conversations contains a total number of 1,574 sentences that make use of finite adverbial clauses, accounting for 5.34% of the total number of sentences in the sub-corpus. On the other end of the continuum, as another example, we duly observe a higher proportion of finite adverbial clauses, that is, 12.81% in terms of sentences and 13.53% in terms of

clauses. It is important to note that this general trend can be observed for all of the different types of adverbial clauses.

Table 4. Types of adverbial clauses across samples of spontaneous and scripted speech

		Spontaneous (29,490)		Scripted (5,793)		Total (35,283)	
		#	%	#	%	#	%
A_{sub}	Sentence	1574	5.34	742	12.81	2316	6.56
	Clause	1757	5.96	784	13.53	2541	7.20
A_{infin}	Sentence	271	0.92	253	4.37	524	1.49
	Clause	279	0.95	260	4.49	539	1.53
A_{ing}	Sentence	190	0.64	161	2.78	351	0.99
	Clause	193	0.65	163	2.81	356	1.01
A_{edp}	Sentence	21	0.07	35	0.60	56	0.16
	Clause	21	0.07	36	0.62	57	0.16
Total	Sentence	2056	6.97	1191	20.56	3247	9.20
	Clause	2250	7.63	1243	21.46	3493	9.89

It is thus reasonable to suggest that within speech the proportion of adverbial clauses increases as a function of degrees of elaboration, formality, and preparedness.

3.4 Types of adverbial clauses across timed and untimed essays

Having established that in speech the proportion of adverbial clauses is largely a function of elaboration or formality or preparedness, we want to do the same for the written samples. We want to argue, on empirical basis, that adverbial clauses not only mark a spoken-written division, that they also mark a continuum between what is spontaneous and what is scripted in speech, and that they also mark a degree of preparedness in writing.

Conveniently, the ICE-GB corpus contains a category coded W1A, which includes 20 texts evenly divided into two sets. Both sets were unpublished essays written by university students. The only difference is that the first set was written within a pre-designated period of time while the second set comprises samples written without the time constraint. If the higher use of adverbial clauses were indeed the result of a higher degree of elaboration or preparedness, then we would observe more uses in the untimed set than in the timed set. This consideration led to a third experiment, whose results are summarised in Table 4.

Table 5**.** Types of adverbial clauses across samples of timed and untimed essays

		Timed (1,057)		Untimed (1,046)		Total (2,103)	
		#	%	#	%	#	%
A_{sub}	Sentence	156	14.76	203	19.41	359	17.07
	Clause	171	16.18	235	22.47	406	19.31
A_{infin}	Sentence	62	5.87	61	5.83	123	5.85
	Clause	65	6.15	64	6.12	129	6.13
A_{ing}	Sentence	59	5.58	51	4.88	110	5.23
	Clause	59	5.58	55	5.26	114	5.42
A_{edp}	Sentence	10	0.94	16	1.53	26	1.23
	Clause	10	0.94	16	1.53	26	1.23
Total	Sentence	287	27.15	331	31.64	618	29.29
	Clause	305	28.86	370	35.37	675	32.09

Again, we duly observed a consistent increase in the proportion of adverbial clauses from one end of the continuum, timed essays, to the other end of the continuum, untimed essays. For instance, we observe that there are 16.18 finite adverbial clauses per 100 sentences for the timed essays. The untimed essays make more uses of finite adverbial clauses, 22.47 per 100 sentences. The same trend can be observed for all of the different types of adverbial clauses, except the infinitival ones. 62 sentences were observed to contain a total of 65 adverbial clauses in timed essays. In the untimed essays, 61 sentences were found to use a total of 64 infinitival adverbial clauses. While the differences are only marginal and can be dismissed as occasional, this group of texts will be examined in a future study for a possible relation between text types and uses of infinitival clauses.

For the purpose of the current study, it can be observed that in the untimed essays as a whole 31.64% of the sentences made use of adverbial clauses, almost 4.5% higher than 27.15% for the timed group. The results thus support the suggestion that within writing the proportion of adverbial clauses indicates different degrees of preparedness in terms of time.

3.5 Discussions

We have thus observed that, in the first place, adverbial clauses mark a division between spoken and written English in the sense that the spoken samples have a lower proportion of adverbial clauses than the written samples. This is true not only for finite adverbial clauses but non-finite ones, including infinitival, present participial and past participial constructions. Secondly, the experiments also produced empirical evidence that the frequency distribution of adverbial

clauses follows a predictable and regular growth curve from spontaneous conversations to scripted public speeches. The same trend can be observed from within the written sample themselves, where the proportion of adverbial clauses in general increase from timed essays to untimed essays. As Figure 2 clearly demonstrates[1], the proportion of adverbial clauses per 100 sentences in ICE-GB consistently increases along a continuum between spontaneous conversations and untimed university essays. What is remarkably surprising is the fact that the occurrence of adverbial clauses in spontaneous conversations accounts for only about 7.5% of the utterances. What is equally surprising is that the occurrence of adverbial clauses in untimed university essays accounts for over 35% of the sentences, over 4.6 times as much as that in speech. The sharp contrast between speech and writing shown in Figure 2 argues strongly against the claims of past studies.

The graph also shows the average proportions of adverbial clauses in the two modes are nicely situated between the two sections within the same continuum. First of all, the average proportion of adverbial clauses in speech is shown in the figure to be between spontaneous conversations and scripted public speeches, suggesting a consistent increase in speech along the 'preparedness' register. In the written section of the continuum, the average proportion of adverbial clauses in writing rests between timed and untimed essays, again suggesting a consistent increase, continuing the trend from the spoken section, along the 'preparedness' register.

While it is evident from Figure 2 that speech and writing demonstrate a vast difference in terms of the use of adverbial clauses, it is clear at the same time that adverbial clauses are not as much a factor of speech vs writing division as a degree of preparedness in discourse. To be exact, it is acceptable to suggest on the basis of empirical evidence that degrees of information elaboration dictate the proportion of adverbial clauses: the more elaborate the sample (defined in terms of preparedness), the more adverbial clauses. The results are thus significantly different from those past studies such as Thompson (1984) and Bibier (1988).

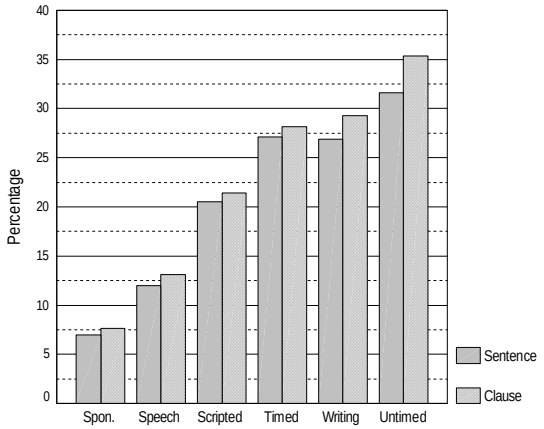


Figure 2**.** The increase of adverbial clauses as a function of degrees of preparedness

4. Conclusion

To conclude, this article first of all described the design, construction and annotation of an interoperable resource, ICE, to enable interoperable linguistic studies. Interoperability was ensured at different key stages of the project through standard annotation schemes and automatic systems that apply these schemes. The article then described a linguistic investigation into the use of adverbial clauses across speech and writing on the basis of ICE-GB that has been grammatically tagged and syntactically parsed. The detailed syntactic annotation of the corpus and manual validation of the analysis ensured that adverbial clauses could be accurately retrieved. Results suggest that, contrary to claims by past studies, the proportion of adverbial clauses is generally lower in speech than in writing. It is also shown that adverbial clauses do not simply mark a division between the spoken and written genres. Empirical evidence also suggests that the proportion of adverbial clauses is also a function of varying degrees of preparedness, which can be independently demonstrated from within the spoken and written genres. It is thus reasonable to postulate that the spoken-written division is perhaps better perceived as a continuum of preparedness, from spontaneous private conversations at one extreme to untimed carefully prepared writing at the other, along which the proportion of adverbial clauses consistently change in a predictable fashion.

It is not yet obvious how to account for the different results regarding the use of adverbial clauses across

[1] The *X* axis in Figure 2 has legends indicating the proportion of adverbial clauses in the following groups of samples in ICE-GB:

- *Spon*: spontaneous conversations
- *Speech*: complete spoken samples
- *Scripted*: scripted broadcast news and talks
- *Timed*: timed university essays
- *Writing*: complete written samples
- *Untimed*: untimed university essays

speech and writing. One possible explanation is accuracy of analysis: past studies largely used hand analysed data or automatically analysed data without manual validation. A second possible explanation may have to do with different definitions of the adverbial clause. Temporal prepositions like *before* and *after* are often complemented by gerundial clauses. Such constructions are analysed as prepositional phrases according to ICE manuals but may have been treated as adverbial clauses in some of the past studies. This possibility demonstrates the necessity for use of standardised terminologies in language resources that are truly interoperable.

Acknowledgement

This work was supported in part by an SRG grant from City University of Hong Kong.

References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Fang, A.C. 1996a. Grammatical tagging and cross-tagset mapping. In S. Greenbaum 1996. pp 110 – 124.

Fang, A.C. 1996b. The Survey Parser: Design and development. In S. Greenbaum 1996. 142 – 160.

Fang, A.C. 2000. From Cases to Rules and Vice Versa: Robust Practical Parsing with Analogy. In *Proceedings of the Sixth International Workshop on Parsing Technologies, 23-25 February 2000, Trento, Italy*. pp 77 – 88.

Fang, A.C. 2008. Measuring a Syntactically Rish Parser with an Evaluation Scheme for Automatic Speech Recognition. In *Proceedings of the First Workshop on Syntactic Annotations for Inter-operable Language Resources*, Hong Kong, 8 January 2008.

Greenbaum, S. 1992. A new corpus of English: ICE. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991*, ed. by Jan Svartvik. Berlin: Mouton de Gruyter. pp. 171 – 179.

Greenbaum, S. 1996. *The International Corpus of English*. Oxford: Oxford University Press.

Greenbaum, S. and Y. Ni. 1996. About the ICE Tagset. In S. Greenbaum 1996. pp 92 – 109.

Thompson, S. 1984. Subordination in Formal and Informal Discourse. In D. Schffrin (ed), *Meaning, Form, and Use in Context: Linguistic Applications*. Washington DC: Georgetown University Press. pp 85 – 94.

Can WordNet and FrameNet be made "interoperable"?

Collin F. Baker
FrameNet Project
International Computer Science Institute
Berkeley, California
collinb@icsi.berkeley.edu

Christian Fellbaum
Cognitive Science Laboratory
Princeton University
Princeton, New Jersey
fellbaum@clarity.princeton.edu

Abstract

FrameNet and WordNet are two lexical databases that are widely used for NLP, often in conjunction. Because of their complementary designs they are obvious candidates for alignment, and an exploratory research project has begun on this topic.[1] We discuss some specific problems that would need to be solved in order to enable interoperability and outline possible solutions.

1 WordNet and FrameNet

1.1 WordNet

WordNet (hereafter WN) is by far the largest freely available lexical database for English and it has become the lexicon of choice for most NLP research; Kilgarriff (2000) notes that "NOT using it requires explanation and justification."[2] Its digital format and network structure, where word senses are interconnected on the basis of their meanings, make WN a tool well-suited for the many applications that require word sense disambiguation. Being the only lexical database of its kind, WN has been used as the standard inventory of word senses for evaluation exercises on word sense disambiguation (Mihalcea and Edmonds, 2004), and measuring semantic distance via links in the hierarchy (for a summary of similarity measures based on WN see (Pedersen et al., 2004; Budanitsky and Hirst, 2006)).

[1]This paper reports on work supported by the National Science Foundation under Grant IIS-0705199.

[2]See http://engr.smu.edu/~rada/wnb/ for a bibliography of WN-based work.

The building blocks of WordNet are synonym sets, known as *synsets*. Each synset is an unordered set of distinct word forms, or *lexemes*. For two lexemes to belong to the same synset it is required that they refer to the same concept and be "cognitively synonymous" (at least on one reading of each) (Cruse, 1986, 270-285). More formally, synset members must be interchangeable in some contexts without altering the truth value of the context; examples of synsets (marked in this paper by curly braces) are {car, automobile} and {shut, close}. Each synset contains, in addition to the synonyms, a definition, or "gloss", and sometimes example sentences. The current version of WordNet (3.0) contains over 117,000 synsets, but many contain only one member.

1.1.1 Lack of Syntax

Apart from a limited number of subcategorization frames for a subset of its verbs, WN contains no syntactic information. It is thus not possible to align WN's semantic classes with a syntactic classification. Moreover, for many but not all verbs, WN gives distinct (though related) senses for causative, inchoative, and middle uses. If we think in terms of their superordinates, inchoatives ultimately link to {undergo_a_change}, while causatives link to a superordinate {cause_to_change} (and have one more participant in the event), and thus merit treatment as separate senses. Middles, on the other hand, actually have a superordinate {be}; lacking a way to analyze this form as built up by a "middle construction", WN is forced to treat them as separate senses, also. Syntactic alternations involving Location and Locatum

arguments for verbs like {spray} and {load} also require the different superordinates {cover} and {put} and are treated as separate senses.

1.2 FrameNet

FrameNet (hereafter FN) is a lexicon of English which is intended to be both human- and machine-readable based on the theory of frame semantics (Fillmore, 1982), which asserts that the meanings of many words are best understood in terms of an entire situation and the participants and props involved in it; the situation is called a **frame**, and the participant roles are called **frame elements (FEs)**. The link between a lemma[3] and a frame is a **lexical unit (LU)**, which is roughly equivalent to a word sense in a conventional dictionary, or to a WN sense.

FrameNet is recognized as providing a high-quality lexical resource that has applications in many NLP tasks, such as word sense disambiguation, information extraction, question answering, and the emerging field of automatic semantic role labeling (ASRL). The FN data, which is freely available for research purposes, has been downloaded more than 400 times, and is being used around the world, including work on FrameNets in other languages.

1.3 Differences between the Resources

These two lexical databases, are in many ways complementary; WN's coverage is very extensive but it has little syntactic information, while FN's information on argument/adjunct realization patterns is very detailed, but only for about 10,000 word senses. The most common criticism regarding FrameNet is that it does not have sufficient coverage; the most common problem concerning WordNet is that some of the sense distinctions are too fine-grained. The major differences between WN and FN are:

(1) WN has separate networks of synsets for nouns, verbs, adjectives, and adverbs, with different types of synset-synset relations for each; FN groups words of all parts of speech in a single frame if they evoke the same type of event[4], and groups almost all of the frames into a single network.

(2) FrameNet concentrates on verbs, event nouns, relational adjectives and prepositions, and does not aim to include all the names of artifacts, natural kinds, etc., which do not evoke interestingly different semantic frames, although a few such frames have been created as experiments (e.g. Natural_features, Clothing). (Cf. Sec. 6 for more discussion.) WN is more than ten times larger, and contains thousands of quite specific nouns.

(3) WN has only very general syntactic frames at the level of synsets, only for some verbs; FN has specific valences for each LU (of whatever part of speech), derived from annotated corpus examples.

(4) FN treats many supposed senses of verbs not as separate senses but only as supports for nouns. E.g. Consider the phrase *place emphasis on*; WN treats this as an instance of placing, but in an abstract location. FN, however, includes *emphasis.n* in the frame Place_weight_on; in this phrase, *emphasis* is treated as the semantic head, occurring with supports such as *give, lay, place, put*, and *with*.

In the remainder of the paper, we discuss some questions related to efforts in aligning the two resources.

2 Comparison with other alignment efforts

Recognizing both the value and the complementary nature of the two resources, a number of researchers have attempted to map between FrameNet and WordNet. We briefly discuss the major efforts and their shortcomings.

Broadly speaking, most semi- or fully automatic mappings are based on assumptions that do not always hold. The first is that WordNet synset membership entails membership in the same frame or frames, i.e., synonyms should be equivalent LUs in a given frame. Second, words and synsets related via WordNet relations are likely to share membership in the same frame.

WordNet's synsets were constructed on the basis of close semantic similarity. Synset members refer to the same concept, a purely semantic criterion. WordNet synset members often have different syntactic properties. Thus, the verbs in the synset *hate, detest* seem to express the same concept. But they are not interchangeable in all syn-

[3] A lemma is the basic form of a word, which may have more than one word forms associated with it. Thus the verb *give* is a lemma, with word forms *give, gives, gave, given, giving*.

[4] Due to space limitations, we will discuss only nouns and verbs in this paper, but adjectives, adverbs and prepositions are treated similarly in FN.

tactic frames. While both verbs can select complements in the form of direct argument and gerunds ("he hates/detests his boss/cleaning his room"), only *hate* can take an infinitive phrase: "He hates/*detests to clean his room."

Some linguists believe that semantic similarity is closely correlated with syntactic similarity. This facilitates children's acquisitions of the meaning and use of verbs Gleitman (1990) and allows speakers to agree on the use of new verbs like *e-mail* and *regift*. Levin's (1993) syntactic examination of thousands of English verbs yields coherent semantic classes and seems to confirm the thesis that syntax and semantics are correlated.

VerbNet (Kipper et al., 2000) represents an effort to connect the syntactically motivated Levin classes with WordNet's purely semantic classes. However, the Levin classes are syntactically and semantically heterogeneous. Some appear to have made their way into the VerbNet classes; also, mappings from Levin classes onto WordNet lexemes appear to have imported more syntactically heterogeneous verbs.[5]

PropBank (Palmer et al., 2005) is an annotated corpus that considers not just individual tokens but, like FrameNet, entire phrases and the relations among their constituents. Specifically, the arguments of verbs are labeled with roles; these are part of the corresponding verb entries in VerbNet. VerbNet's role labels differ significantly from the ones assigned in FrameNet. Some have semantic content (e.g., Location), others are mainly based on "deep" syntactic position (e.g., Arg0, Arg1). While the syntactic arguments can be given semantic contents in many cases for specific verbs, a systematic assignment is not apparent. FrameNet distinguishes many more semantic roles than VerbNet, necessitating many-to-one mappings in many cases, but the mapping is likely to differ across frames.

Moreover, the assignment of verbs to frames is not identical in VerbNet and FrameNet. Thus, VerbNet analyzes the *speak* and *talk* as "transfer" verbs, treating the "Topic" argument as a transferred entity. The notion of "transfer" and the frames associated with this concept thus differ across the two resources, as does the semantic content and the syntactic behavior of the associated arguments or Frame Elements. A mapping cannot not be made on the basis of the names of the frames or of the FEs, nor of the lexical units involved in a given frame.

Other efforts to automatically map FrameNet and WordNet have been made. Shi and Mihalcea (2005) integrate WordNet, FrameNet, and VerbNet for semantic parsing. They, too, assume that verbs belonging to the same Levin class are likely to share the same Frame. Shi and Mihalcea map VerbNet to WordNet by linking the roles in VerbNet to top level synsets in WordNet and extend them to all hyponyms. This step depends on the categorization in WordNet being flawless, but WordNet has many odd category assignments (for example, "fictional_animals" are kinds of "animals"), which are imported into Shi and Mihalcea's FrameNet-WordNet mappings via VerbNet.

The goal of Burchardt et al. (2005) is to increase the number of frames in FrameNet; this is done via a WordNet "detour." Already annotated frames, where the LUs are linked to WN synsets, serve as training data for a learning system. Frame assignment is done not on the basis of a single target word but also includes its synonyms, hyponyms, and antonyms. All frames evoked by all lexemes related to the target words are determined and evaluated. In this way, frames can be assigned to words that are not yet covered by FrameNet.

Burchardt et al.'s method, too, rests crucially on the assumption that the WordNet similarity among lexemes is reflected in shared frame membership. However, this is often not the case. As we already saw, synsets are semantically close but their different syntactic properties frequently reveal subtle semantic differences. Hyponyms may be quite different from their superordinates not only because of miscategorization but also for lack of a more appropriate lexicalization for an arguably existing concept. Oltramari (2006) elaborates Burchardt et al.'s method with statistical analysis of the frame distribution among the hyponyms.

Chow and Webster (2007) undertake a mapping

[5] For example, VerbNet lists both *hate* and *detest* as members of the "admire" class (31.2-1), along with *despise, disdain, dislike, enjoy, fear, like, love, regret*; many members of this class show the same syntactic restriction as *detest* and cannot select for an infinitival complement (*She despises/fears to work). It is not immediately clear whether this syntactic distinction can be related to a meaning distinction among the verbs, but the example shows at least that semantic similarity — as reflected in shared synset membership — is not necessarily accompanied by complete syntactic similarity.

between FrameNet and WordNet via SUMO(Niles and Pease, 2003). Starting with Shi and Mihalcea's mappings of FrameNet LUs to WordNet senses, they add all synset members to the LUs plus all synsets related by all WordNet relations. In this way they obtain what they call a domain, based on the assumption that all synsets that link to one another in WN are in the same semantic domain. To clean the output of this process, Chow and Webster refer to domains in SUMO and existing SUMO concept-WordNet synset mappings. Like other mappings, Chow and Webster assume that WordNet synset members are equivalent to LUs in a given frame.

Ide (2006) presents a promising approach, where the direction of the mapping is from LUs in FrameNet frames to WordNet synset members. Ide uses Pedersen's similarity measures to map specific WordNet senses with FrameNet LUs. All LUs of a given frame are submitted to the similarity measures–among which the Lesk measure seems to be the best–and a similarity rating is obtained. The most similar words are assumed to be mutually disambiguating and the most appropriate WordNet sense can be determined. Ide's method does not rely on WordNet relations, which are not intended to capture syntactic similarity or shared frame membership.

3 Frames and Synsets as sets of word senses

The difficulties enumerated above should not lead to the conclusion that all efforts to relate WN synsets (SSs) to FN frames are fruitless. Indeed, at least one of the results of the study is expected to be in the form of a database recording relations between WN SSs and FN frames. The first step is to manually identify corresponding FN LUs and WN synsets.

One of the advantages of involving both FrameNet and WordNet in this effort is that, where the results suggest that the treatment in one of the resources is right and the other is wrong, the wrong one can be corrected. This is exploratory research, but certain types of (mis)alignment and their resolution can already be foreseen:

1. In the very unlikely case that a SS and a frame contain exactly the same set of lexemes, their correspondence will simply be recorded.

2. In the more common case in which all the words in a SS are a subset of those in the frame, or all the words in a frame are a subset of those in the SS, this fact will also be recorded in the DB.

3. In case two SSs are subsets of the LUs of one frame, we will record this and note it as a possible candidate for collapsing the SSs, respectively. [6]

All of these comparisons will be made first for whole sets, and then on the basis of partial matches, with experimentation as to what criteria produce the best input to the human decision process.

Given the much greater size of the WN vocabulary, we expect that one of the most common changes will be to add LUs to a frame from the related synset(s). In some cases, the added LUs will be relatively rare, as WN deliberately includes many rare words[7]. If they are to be added to FrameNet, they should also be documented by annotated examples; we will consider carefully how to extract enough examples for such words, what corpora might contain them or how to find them on the web, etc. In many cases, it may be necessary to mark them as being insufficiently attested, as certain words in FN are now.

4 FrameNet and WordNet as Directed Acyclic Graphs

As noted above, both WN and FN contain a variety of links between their SSs and frames respectively, many of which will be useful in the project of aligning the resources.

4.1 Frame relations in FrameNet

FrameNet currently contains 887 frames, connected by 1,298 frame-to-frame relations, of seven types, comprising three major groups.[8]

[6] The reverse, when the LUs of two frames are both subsets of one SS, is rarer, but we will also consider collapsing the frames in that case.

[7] WN includes both the most frequent words and many quite rare ones, such as *frore* 'frigid, frozen' and *amercement* 'fine, penalty'.

[8] There are actually 712 other frame relations in the database which are not relevant to this discussion.

Count	Name	Group
538	Inheritance	
69	Perspective_on	Generalizations
465	Using	
105	Subframe	Complex events
66	Precedes	
42	Causative of	Aktionsart
13	Inchoative of	

The first group of three relations has to do with generalization/specialization, which means that the situation denoted by the child frame must be a subtype of the situation denoted by the parent frame. They account for more than 2/3 of the instances, and join the frames into three hierarchies for events, relations, and states, which are then "cross-linked" by the other two groups of relations, having to do with complex events and aktionsart. Together, these relations connect almost all of the frames in FN into a single graph, though there are a few "islands" that are not part of the main group.

Inheritance is the strictest of these generalization relations. For example, the **Intentionally_affect** frame depicts a very general type of event event in which " An Agent causes a Patient to be affected,sometimes by a particular Means or by use of an Instrument." This is inherited by many frames, among them the **Inhibit_movement** frame, defined thus: "An Agent restricts the movement of a Theme to within the vicinity of the Holding_location, despite the Theme's desire, plan, or tendency towards motion. . . [Some] LUs may be used to describe punishment situations when the Theme is a sentient entity." As noted in the last sentence, some of the LUs in this frame, such as *imprison*.v, occur only with sentient Themes, (in the annotated examples, all are human). Finally, the **Imprisonment** frame, which inherits from **Inhibit_movement**, is defined in strictly legal terms: "The Authorities put a Prisoner in Prison as punishment for an Offense", and most of the LUs, such as *incarcerate*.v and *jail*.v can only be used of literal, legal confinement of people.[9]

[9]However, the verb *imprison* is listed in both the **Inhibit_movement** frame and the **Imprisonment** frame; there are no annotated examples in the more specific frame, yet almost all the sentences annotated in **Inhibit_movement** actually fit the more specific definition, suggesting that they should be moved to the lower frame.

In FrameNet practice, the decision as to what LUs belong in a frame is inextricably connected with the decision as to what FEs are needed for it. Thus the **Imprisonment** frame adds the FE OFFENSE to the set inherited from **Inhibit_movement**; cf. the different interpretations of Ex. (1-a) and (1-b).

(1) a. He was confined [$_{\text{REASON}}$ for observation]. (**Inhibit_movement**)
b. He was jailed [$_{\text{OFFENSE}}$ for armed robbery]. (**Imprisonment**)

The next generalization relation, the **Perspective_on** relation, has to do with the profiling of different parts of an event depending on which participant's perspective is taken. For example, the **Employment scenario** is different from the points of view of the EMPLOYER and the EMPLOYEE; the stage of the scenario which can be abstractly described as **Employment_start** is called **Hiring** from the EMPLOYER's perspective, and **Get_a_job** from the EMPLOYEE's perspective. But **Employment_start** is a non-lexical frame, created solely to build the hierarchy linking the more concrete lexical frames with a higher-level frame **Employment** which houses the unperspectivized LU *employment*.

Finally, the **Using** relation is the vaguest of the three. In these cases, the parent frame provides a general background for understanding the event; often the child has an inheritance link to a more generic scenario, which is usually associated with a particular syntactic argument structure.

4.2 Synset links in WordNet

The synsets are linked to one another via a small number of binary relations that differ for each of the four syntactic categories covered by WordNet: nouns, adjectives, verbs, adverbs.[10] Noun SSs are interlinked by means of **hyponymy** (the super-/subordinate or IS-A relation), as exemplified by the pair [poodle]-[dog], and **meronymy** (the part-whole or HAS-A relation), as in [tire]-[car] (Miller, 1989). Verb SSs are connected by a variety of lexical entailment pointers (Fellbaum, 1998) that express concepts such as manner elaborations ([walk]-[limp]), temporal relations ([compete]-[win]), and causal re-

[10]WordNet's adverb component is small and will not be considered further in this paper.

lations ([show]-[see]). These links between synsets structure the noun and verb lexicons into separate hierarchies, with the noun hierarchies being considerably deeper.

Most relations in WN connect SSs whose members belong to the same lexical category. More recently, cross-POS links were added, but only between words that are both semantically and morphologically related (Fellbaum and Miller, 2003). For example, the noun *director* has two senses {manager} and {theater director}, which are linked two senses of the verb *direct*, {be in charge of} and {direct actors}, respectively.

5 Alignment using parallel relations

5.1 Within Part-of-speech

The most likely candidates for aligning the two resources would be hypernymy/hyponymy links between verb SSs, which will often correspond to generalization relations in FrameNet. For example, the FN inheritance relation between **Inhibit_movement** and **Imprisonment** discussed above is paralleled by WN links between hypernym *detain* and *incarcerate*.

Some causative relations are also represented in both resources; e.g. WN has two senses of *change*, 'cause to change' and 'undergo change', with a Cause_to link between them; FN has *change*.v in both the **Cause_to_change** and **Undergo_change** frames, with a Causative_of relation between them. This seems to be an area in which FN is more complete than WN; e.g. *kill* as a causative of *die* is represented in FN but not in WN.

The Meronym/Holonym relation, on the other hand, is not really represented in FN, but there are thousands of such relations in WN, telling us that a steering wheel is a part of an automobile, that lakes are full of water, etc. FN has no plans to add this sort of information.

5.2 Across Part-of-speech

5.2.1 WN Derivation links and FN Frames

The cross-POS links in principle should accord well with the fact that FN frames include all parts of speech. Thus, the WN derivational links between *leader*.n 'a person who rules or guides..." and *lead*.v 'be in charge of' are reflected in the fact that both the noun and the verb are in the FN **Leadership** frame; i.e. the semantic relation that hold among the relevant senses of *director* and *direct* also holds among their synonyms. But, although *director*.n is also in **Leadership**, *direct*.v is not; this is purely an omission in FN, and the WN links will serve to suggest this additional LU for the frame.

WN morphosemantic links have no semantic content; thus, WN does not tell us that "director" is the Actor in a "directing" event. [11] By contrast, the FN frame **Leadership** contains both the verbs *lead* and *manage* and the nouns *leader*, and *manager*. Also, these nouns are marked with a special semantic type indicating that they refer to agents rather than to events, even though the frame as a whole represents a type of event. Thus the FN frames serve to link semantically across the separate WN hierarchies for different POS, even for morphologically unrelated words.

5.2.2 FN Subevents ("Subframes") and WN entailment relations

WN encodes an *entailment* relation among verb synsets. These are really several distinct relations, but they all express the logical necessity of one event given another. Examples are *buy* and *pay* (a buying event necessarily entails a paying event), *divorce* and *marry* (to divorce you must necessarily have married), and *breathe* and *live* (one cannot live without breathing). The verbs *marry* and *divorce* (along with the nouns *marriage* and *divorce*) are annotated in FN in the **Forming_relationships** frame, but the entailment between them is not (currently) represented. On the other hand, the first of these is well represented in FN: this sense of *pay* is an LU in the **Commerce_pay** frame, which is the BUYER's Perspective_on the **Commerce_money_transfer** frame, which is a Subframe of the **Commercial_transaction** frame; this sense of *buy* is in the **Commerce_buy** frame, which is the BUYER's Perspective_on the **Commerce_goods_transfer**, which is the other Subframe of the **Commercial_transaction** frame. In other words, the completion of a commercial transaction

[11] In a current collaboration with Boeing, the Princeton WN team is addressing this shortcoming and adding semantic labels to the morphosemantic links (Fellbaum et al., 2007). Labels including Actor, Instrument, Location, etc. are attached to the links characterizing the relation among the nouns and verbs.

requires both types of transfer, and from the buyer's point of view, these are expressed as buying the goods and paying the money.

6 "Covering" Frames and WN Subtrees

A special relationship exists between certain frames and certain subtrees of the WN N hierarchy. These are places where the FN staff has long been aware that there is a large group of nouns which, (1) from a frame-semantic point of view behave rather similarly and (2) fall into a single subtree in WN rather logically. Examples include animals, plants, landforms, bodies of water, etc.

In each case, there is a FN frame which contains all the FEs needed for the entire class of nouns. And in each class, there is a detailed hierarchy in WN, which does a good job of representing the ontological distinctions. In such domains, the obvious choice is to let each resource contribute the information which it can to the combined database.

For example, consider the sentence:

(2) The Yapok is a nocturnal, aquatic, South American marsupial.

What we would like to have happen is for FrameNet to supply the linking structures, to explain how *nocturnal*, *aquatic*, and *South American* are related to *marsupial*, and for WordNet to supply the Linnaean information about marsupials, that a marsupial IS-A metatherian, IS-A mammal, IS-A vertebrate, etc.[12] In fact, the IS-A hierarchy in WordNet is already pretty complete and accurate for animals and plants. FrameNet has not created a similar hierarchy, nor does it seem necessary to do so, because the frame elements and the syntactic positions in which they occur will be essentially identical for broad swathes of the animal kingdom[13], and the hierarchy itself would largely duplicate that of WordNet. What FrameNet **could** usefully indicate is that animals are often described in terms of their range, habitat, time of greatest activity, etc., which we could regard as FEs of a general **Animal** frame, annotating Ex. (2) something like this:

(3) The Yapok is a [$_{\text{ACTIVITYTIME}}$ nocturnal], [$_{\text{HABITAT}}$ aquatic], [$_{\text{RANGE}}$ South American] marsupial.

Deciding exactly how this can best be accomplished in operation is one of the objectives of the current research. Obviously, FN could create one huge frame called **Animal**, and list thousands of animal names in it, (which would greatly boost the number of LUs) but even that would be a duplication of WordNet. Another approach would involve a link to WN, whereby finding a specific animal name in a sentence will trigger a search of the WN hypernym tree; when the *animal* synset is reached, this will activate the FN **Animal** frame, allowing the FEs to be used to annotate the sentence.

7 Towards a more dynamic view of word meaning

Creating lexical resources forces the lexicographer to commit to an inventory of enumerable and discrete senses (but see Pustejovsky (1995) for an alternative).[14] But a cursory comparison of different dictionaries shows that even professional lexicographers do not agree on how the different meanings of polysemous word should be divided up. Unsurprisingly, both people and automatic systems, when asked to assign tokens in a text to the appropriate senses in dictionaries, find the task difficult and do not agree among themselves (Fellbaum et al., 1997). This indicates that word senses are not discrete but have fuzzy, overlapping boundaries, and in many cases are structured around prototypical uses. For example, WordNet distinguishes several senses of 'play' that all involve an artistic performance but vary in the semantics of their arguments: *play the flute/Bach/ a minuet/Carnegie Hall*. But clearly there is a prototypical sense that subsumes all of these. Fuzziness is also apparent in the gradation in degree of lexicalization of metaphorical uses, making it very difficult in practice to decide what is an established sense and what is an ad-hoc use.

We do not expect to be able to resolve all the con-

[12] This kind of information is also contained in the definitions ("glosses") of the SS. The words in the glosses have been manually annotated against the context-appropriate WN senses.

[13] Of course, the variety of the biological kingdoms means that some differences will occur: e.g., whatever FE *anaerobic* belongs in, it will not apply to vertebrates.

[14] Note also that Levin's verb classification avoids reference to senses and merely assigns a given word form to syntactically motivated classes.

tested issues in lexical semantics in this study, but we hope at least to point the way toward a better utilization of the strengths of these two major lexical resources.

References

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schrder, and Petra Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*, pages 408–421. Peter Lang, Frankfurt.

Ian Chow and J. Webster. 2007. Integration of linguistic resources for verb classification: FrameNet, WordNet, VerbNet, and suggested upper merged ontology. In *Proceedings of CICLing*, pages 1–11.

David Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

Christiane Fellbaum and George A. Miller. 2003. Morphosemantic links in WordNet. *Traitement Automatique des Langues*, 44:69–80.

Christiane Fellbaum, Joachim Grabowski, and Shari Landes. 1997. Analysis of a hand-tagging task. In Marc Light, editor, *Tagging Text with Lexical Semantics: Why, What and How?* Special Interest Group on the Lexicon, Association for Computational Linguistics.

Christiane Fellbaum, A. Osherson, and P.E. Clark. 2007. Putting semantics into WordNet's "morphosemantic" links. In *Proceedings of the Third Language and Technology Conference*, Poznan, Poland.

Christiane Fellbaum. 1998. A semantic network of English verbs. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Lila Gleitman. 1990. The structural sources of verb meaning. *Language Acquistion*, 1:3–55.

Nancy Ide. 2006. Making senses: Bootstrapping sense-tagged lists of semantically-related words. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text*, Lecture Notes in Computer Science. Springer.

Adam Kilgarriff. 2000. Review of wordnet: An electronic lexical database. *Language*, 76:706–708.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence*, Austin, TX. AAAI-2000.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Rada Mihalcea and Phil Edmonds, editors. 2004. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics.

George A. Miller. 1989. Nouns in WordNet. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, Las Vegas, Nevada.

Alessandro Oltramari. 2006. Lexipass methodology: a conceptual path from frames to senses and back. In *Proceedings of LREC 2006*, Genoa, Italy.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Lingustics*, 31(1):71–106, March.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 38–41, Boston. NAACL-2004.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Proceedings of Cicling*, Mexico. Download at http://www.cs.unt.edu/řada/downloads.html.

Challenges for a Global WordNet

Christiane Fellbaum
Princeton University
Princeton, NJ (USA)
fellbaum@princeton.edu

Piek Vossen
Free University of Amsterdam
Amsterdam, The Netherlands
Piek.Vossen@hum.uva.nl

Abstract

We describe the overall design and principal components of a Global WordNet, whose principal goal is to provide an environment for the interoperability of multilingual present and future wordnets that is beneficial for linguistic-lexicological studies and NLP applications. Multilingual wordnets will capture the inventory of lexicalized concepts in wordnets of many typologically and genetically unrelated languages. The list of lexemes to be included is not exlusively determined by linguistic rules based on economy but also by frequency, salience, and cultural significance. Each wordnet is linked to a shared formal, language-independent ontology that can represent their lexical inventories in a language-independent way and that is amenable to reasoning and inferencing. We discuss the division of labor between the wordnets and the ontology with respect to meaning representation and relations.

1 Introduction

There is a pressing need for coordinating the many wordnet efforts (currently well over forty), as the majority are developed independently of one another. Some simply translate the Princeton wordnet synsets into the target language; others work monolingually and map their networks onto existing wordnets. In both cases, solutions for mismatches are often formulated from a language-specific rather than a global perspective, which hampers potential interoperability among the databases.

The Global WordNet project builds on two predecessors: the Princeton WordNet and EuroWordNet. Each one was limited in its scope and goals, but provided experience that is valuable for the effort outlined here.

1.1 Background: Princeton WordNet

Princeton WordNet (Miller, 1995; Fellbaum, 1998) was an experiment in digital lexicography designed to test theories of human semantic memory. At the time it did not occur to its creators that the result would be of interest to computational linguistics, serve as a tool in NLP applications, and be replicated in other languages; thus the task seemed fairly straightforward. English lexemes (including some compounds, phrasal verbs, and idioms) were extracted from traditional lexical resources and texts or suggested by users. They were then grouped and arranged according to a few well-established semantic relations (hyponymy, meronymy, antonymy) or variations thereof (troponymy). WordNet resembles a thesaurus, but, in contrast to traditional thesauruses like Roget's, the relations among its words and synsets are explicitly labeled and few in number. Moreover, WordNet's links among words representing different lexical classes are limited.

1.2 EuroWordNet

EuroWordNet in the mid-1990s was the first effort to replicate the Princeton work in other languages with the goal of interconnecting the different wordnets. (Vossen, 1998) describes the challenges inherent in

mapping the lexicons of four (later eight) languages.

EuroWordnet had to face the problem that many specific lexicalizations are not universal, even in closely related languages. While this can be expected at the leaf nodes of a "tree," other cases seem surprising, such as the lack of a Dutch words for such top-level concepts as *container* (an object used to hold things) and *artifact* (a man-made object). Words that are not lexicalized crosslinguistically result in mismatched super-/subordinate relations and hierarchies of different depths. This may falsely suggest semantic differences among equivalent lexemes whose position in their respective local networks are not identical: WordNet is built on the assumption that the meanings of words are in part defined by their position in the network, and many applications requiring word sense disambiguation measure the semantic relatedness of words in terms of distance in the "trees."

1.3 The Interlingual Index

The EuroWordNet solution to such lexical mismatches was to create an Interlingual Index (ILI), a list of all concepts that were lexicalized in at least one of the EuroWordNet languages. The basis for the ILI was the Princeton WordNet; each language could add an entry for its words that have no equivalent in English. Unlike the wordnets, the ILI is a flat list and, unlike an ontology, is not structured by means of relations. ILI entries ("records") merely function to connect equivalent words and synsets in the different languages.

Equivalence relations between the synsets in different languages and Princeton WordNet are made explicit in the ILI. Each synset in the language-specific wordnet has at least one equivalence relation to an entry in the ILI. Thus, synsets linked to the same entry in the ILI can directly map the corresponding synsets and words, allowing for a variety of crosslinguistic applications.

In EuroWordNet, all semantic and lexical relations are confined to the wordnets of the individual languages.[1] Encoding relations in the individual wordnets precludes mismatched hierarchies. Moreover, it avoids positing a set of universal relations that are both necessarily and sufficient for the construction of semantic networks in all languages. As wordnets are constructed in more typologically and genetically unrelated languages, they demand additional, language-specific semantic relations.

For example, many — but not all — European languages regularly and productively mark gender in profession and "role" nouns (English does so only in some cases, like *poet/poetess*). If one wants to encode the relation between the members of such pairs, a "gender" relation is needed for these languages, but this relation is far from universal and is absent in, e.g., Japanese. Other examples of language-specific relations is Aspect, the systematic lexical encoding of perfective and imperfective verb forms in languages like Slavic. Arguably, this semantic distinction could be relegated to the grammar rather than the lexicon. We will return to a discussion of the lexis/grammar distinction in wordnets later.

2 What belongs in the lexicons of the Global WordNet?

The central goal of developing and linking crosslinguistic wordnets is to enable the mapping of equivalent lexemes. A collection of a significant number of salient and frequent lexicalized concepts in a broad range of genetically and typologically unrelated languages would be valuable for lexicological and linguistic studies, such as large-scale comparisons of language-specific and universal lexicalization patterns. For NLP purposes, interlinked and interoperable wordnets carry tremendous potential.

But clearly, not all concepts need to be included and the shared coverage must be carefully determined.[2]

For lack of an alternative, many wordnets, including those for Arabic (Rodriguez et al., 2008) and Zulu (LeRoux et al., 2008), start with the inventory of Common Base Concepts originally identified for the EuroWordNet project. But the Base Concepts were defined from a European perspective and they are therefore limited and most likely subject to biases. Consequently, words and concepts that are

[1] EuroWordNet includes a separate, hierarchically structured Top Ontology, where high level concepts like *Location* reside, as well as Domain Ontologies.

[2] For most uses of wordnets, many of the highly specific concepts found in the Princeton WordNet (insects, bacteria, chemical compounds, etc.) are unlikely to be of relevance; a more appropriate place would be in domain-specific term banks and ontologies that cold be linked to a general domain wordnet.

frequent and important to speakers of other, non-Western languages, may be overlooked when wordnets are built around the Base Concepts, even when the core is extended with hyponyms and meronyms.

The Global WordNet attempts to take a broad perspective and include lexemes that are salient for speakers of each language and cultural background. To meet this goal, different and sometimes conflicting criteria of inclusion must be considered.

2.1 Language- and culture-specific words and concepts

Representing the core lexicon of a language in a crosslinguistic system means that highly language-specific concepts will not have a counterpart in other languages and can only be mapped to an entry in the ontology. Examples are Japanese *wabi sabi* (aesthetics based on beauty, imperfection, and impermanence), Dutch *klunen* (travel on skates overland between canals), and German *gönnen* (not begrudge somebody something).

A lack of crosslinguistic relevance does not preclude such words from inclusion in the Global WordNet, which aims to cover not only the words and concepts that are shared among languages but also those that are specific to a single language or a subset of wordnets.

2.2 New words and meanings

Dictionary editors must decide with each new editions which new words and meanings to add. Their criteria reflect those of the Global WordNet: relevance and frequency. Examples of words that have recently been assumed in American dictionaries include *sandwich generation*, *helicopter mother*, and *speed dating*. These express culturally salient, current concepts. Loanwords like *manga*, *Bollywood* and *qigong* enter the lexicons of languages other than those of their origin. Familiar words can assume new meanings to denote newly salient concepts, such as *rendition* (referring to the practice of secretly flying foreign terrorist suspects for interrogation to countries with few regulations for the humane treatment of prisoners). One might argue that such lexemes are transient and likely to disappear again, or that they are relevant to a substratum of speakers only. But since new words and meanings reflect the current lexical inventory of a language and are partuclarly useful for NLP applications, the Global WordNet aims to include them.

2.3 The phrasal lexicon

The phrasal lexicon is the repository of multi-word expressions (MWUs) including idioms like *hit the ceiling, push up daisies, not by a long shot*. These "long words" are arguably part of the lexicon as they are strings associated with meanings that must be learned (Fellbaum, 2007), even though their syntax and lexical make-up is not as fixed as often assumed (Fellbaum and Stathi, 2006). Although some idioms are found crosslingually, many are specific to a given language and may mirror situations, attitudes and behaviors unique to a culture.[3] Idioms are randomly distributed across the lexicon and do not easily pattern in terms of paradigmatic relations. Some have simplex verb synonyms (*kick the bucket-die*).

Constructions like *what is X doing Y?* (Kay and Fillmore, 1999) represent another type of MWU form-meaning pair. They tend to be syntactically idiosyncratic and, like idioms, they are semantically noncompositional. Unlike idioms, they are not lexically pre-filled.

MWUs are frequent and constitute a large and important part of the lexicon. Moreover, their computational treatment is vital to NLP applications. The challenges they pose for inclusion in the Global WordNet — such as their representation in the formal ontology — must be met.

3 The boundary between lexis and grammar

Non-compositional noun phrases (*garage sale, oil skin, high tea*) including metonymic phrases (*Green Beret, Blue Helmet*) clearly belong into the lexicon, which is traditionally considered the repository of all that must be learned and cannot be generated or interpreted via regular and productive rules. But what about compounds like *garage door/house door/church door...*, *deerskin/goatskin/calfskin/buckskin...*, and *mint tea/chamomile tea/pekoe tea...*? Strictly linguistic — and economical — criteria for inclusion

[3](Fellbaum, 2007) argues that many idioms serve as pre-encoded messages appropriate to frequently occurring situations. Speakers need not compose such salient messages "from scratch" each time.

in the Global WordNet would not admit these fully compositional compounds. The same could be said for complex verbs (*shoot dead, breathe deep*) that follow regular patterns of word formation.

The Global WordNet is not subject to the space constraints of a paper dictionary. Instead, its criterion for inclusion are frequency and currency (based on corpus data); lexemes that express salient and culturally important concepts and that characterize a community of speakers will be included.

For NLP applications, it is obviously advantageous to to "load" the lexicon, as identifying and processing MWUs automatically is a significant challenge. Phrases are generally easier to process as units that can be looked up in the lexical database than it is to recognize and re-assemble them.[4]

These considerations may override purely linguistic criteria for including lexemes in the Global WordNet.

4 Ontology

The experience with EuroWordnet and the ILI showed that a language-independent ontology that serves as the hub for all lexicons seems like a good idea for the interoperability of wordnets. It is important to clarify the respective role and contents of the lexicons (wordnets) and the ontology and to ensure that they are coordinated and complementary.

4.1 Lexical vs. conceptual ontologies

Princeton WordNet is sometimes called a (lexical) ontology, although its creators did not have in mind a philosophical construct. In fact, WordNet merely attenpts to map the lexicon into a network organized by means of relations; these are familiar from ontology and often implicit in standard lexicographic definitions. WordNet does not represent an ontology separate from the lexicon, but this separation will be deliberately implemented in the Global WordNet.

The lexicon can be defined as the mappings of words onto concepts. When a lexicon is structured like WordNet, it can reveal whether the mapping is arbitrary or follows certain patterns and principles according to which concepts get labeled with a word. Nevertheless, the concept-word mappings of any given language are to some extent accidental; existing words do not fully reflect the inventory of concepts that is universally available. That inventory can be represented in an ILI or any ontology that is independent from natural language.

A conceptual ontology is an artifact, designed by philosophers who attempt to categorize, define, and structure concepts. It must be able to accommodate all linguistic expressions, regardless of the language in which they are labeled. A large-scale, multilingual ontology can show whether there is universal core of lexicalized concepts. Moreover, the inventory of primitives with which terms are defined in a large interlingual ontology can be fruitfully compared against linguists' proposed universal inventory of semantic components (Wierzbicka, 1985).

4.2 Lexical Gaps

Being a lexical resource, Princeton WordNet tries to avoid the inclusion of non-lexicalized concepts. However, some artificial nodes, such as *wheeled vehicle* were added because they serve to nicely divide intuitively plausible subclasses of synsets from one another and result in "cleaner" hierarchies. For verbs, (Fellbaum and Kegl, 1989) argue for the existence of lexical gaps on syntactic grounds, showing that different syntactic behavior of subclasses indicate a higher-level semantic divison that happens not to be lexicalized in English but that is clearly observed by speakers.

Moreover, the builders of Princeton WordNet justified the inclusions of such nodes on the grounds that the concepts they stand for are often lexicalized in other languages and their absence in English seemed accidental — a lexical gap — rather than motivated. The more languages one considers, the faster the number of lexical gaps grows in all the languages that are being compared. A case in point are kinship terms, which vary widely across languages (Kroeber, 1917). For example, Arabic distinguishes twelve relations that are all subsumed under the single, underspecified English word *cousin*. If one represented each of these senses in the English WordNet as a lexical gap, the kinship hierarchy would be severely distorted.

Each lexicon should contain entries only for those concepts that are lexicalized. The interlingual ontology will not only reveal the universality of proposed

[4] And statistics-based recognition of phrases works at best for high-frequency phrases and very large data sets.

gaps, but also — given that it assumes much of the structuring of concept — make it unnecessary to introduce artificial nodes in the wordnets for structural reasons. Gaps that may be merely motivated by the lexicons of individual languages will not be part of the ontology.

Nevertheless, ontologies mirror wordnets in that their high-level concepts do not have natural language equivalents but serve to introduce structure. Thus, SUMO (Niles and Pease, 2003) contains terms like StateORProvince and ArtificialSatellite. The Global WordNet effort will serve to clarify the distinctions between lexicon and ontology as well as the division of labor between the two.

5 Requirements of an ontology

What is required of an ontology that serves as an interlingua to many different languages? First, its contents should not be linguistic in nature, but formal, with concepts represented as logical expressions. Examples of formal ontologies that have been mapped to WordNet are SUMO (Niles and Pease, 2003) and DOLCE (Gangemi et al., 2002). A formal ontology will be amenable to reasoning and inferencing.

Second, the lexical databases and the ontology must complement one another. In EuroWordNet, all relations are confined to the wordnets, and the ILI is a flat list. But in the Global WordNet, some relations will reside in the ontology, as they are required for the formal manipulation of concepts.

5.1 Why a formal ontology?

(Niles and Pease, 2003) point out that in a lexicon, the meaning of words relies on a human interpretation, rather than on a precise mathematical specification. (Indeed, a quick comparison across dictionaries shows that lexicographers come up with different, though not contradictory, definitions.) (Niles and Pease, 2003) imply that a purely linguistic representation of a word is not sufficient to enable a correct crosslinguistic mappings, whereas a formal definition can function as a kind of universal language and enable creators of new wordnets to verify cross-language links by testing them against formal, logical definitions rather than WordNet's definitions.

But not all formal ontologies represent even straightforward concepts like *inside* in the same manner, so a formal representation may be somewhat subjective as well. Moreover, wordnet creators can check equivalence of synsets and words in different languages in terms of their position in the network, i.e., in terms of their relations to other words and synsets; additionally, similar definitions might give a richer measure of equivalence than a logical formula, which tends to be sparse. The objectivity and preciseness of ontological encoding remains a subject for further exploration in the Global WordNet effort.

While the information about a given concept might be more richly represented in a lexical database, a formal ontology holds potential for reasoning. In particular, a shared crosslingual ontology will enable reasoning and inferencing across languages. The meaning of the terms in the ontology can be tested to a large extent for consistency an automated theorem prover, so that the ontologist need not rely completely on human inspection and judgement. For example, WordNet includes the word *earlier,* but it does not include formal axioms that explain precisely to a computer what *earlier* means. Neither the relations nor the definitions in WordNet would allow a computer to assert that the end of one event is before the start of another if one event is earlier than the other.

6 Representation of meaning in lexicons and ontology

In contrast to a lexicon — a natural phenomeon — - an ontology is an engineered product — an artifact. Whereas a lexical ontology relies on natural language definitions to express the meaning of a word or concept, in a formal ontology meanings are represented by axioms, mathematical statements using first order logic. The names of the terms could be replaced by arbitrary unique character strings and their meaning would still be the same. For more discussion see (Pease and Fellbaum, 2008).

7 Relations in the lexicon and in the ontology

Given that we are striving for an economical and efficient divison of labor between the wordnets and the ontology, the question arises as to where the

relations among words and concepts should reside. Whereas the ILI in EuroWordNet was flat and unstructured and left the assignment of all relations to the individual wordnets, the ontology for the Global WordNet needs to structured as this is required to make it amenable to formal manipulation. But which relations belong into the ontology and which belong into the individual wordnets?

7.1 Lexical vs. conceptual relations

An important distinction is that betwen lexical (word-to-word) and conceptual (concept-to-concept) relations. In the Princeton WordNet, lexical relations include synonymy, antonymy, and morphosemantic relations; these hold among specific words, i.e., synset members. All other relations (hyponymy, meronymy, troponymy, entailment) are conceptual and link entire synsets.

Because an ontology represents meanings formally rather than linguistically and the name for a term is arbitrary, multiple linguistic labels that the language uses to refer to this term (synonymy) is irrelevant to the ontology. (Pease and Fellbaum, 2008) argue further that antonymy does not belong into a formal ontology a it is a purely lexical relation. Other linguistic relations that are likely to be included in the wordnets rather than the ontology include perspective, register and gender.

(Pease and Fellbaum, 2008) argue against including the hundreds of relations that are found in a formally specified logical theory like SUMO, such as *subAttribute*, in the lexical databases, as this would require relating relatively informal linguistic notions with more formal ontological relations. By keeping ontological relations in the formal ontology, and linguistic relations in the lexicon, one can avoid merging two different levels of analysis and yet still capture the information that is needed about both formal concepts and linguistic tokens. A formal ontology such as SUMO also contains formal rules that specify complex relations that cannot be captured explicitly as simple links in a graph.

These formal properties enable reasoning and inferencing for automatic systems that can access and manipulate SUMO much in the same way that humans could use a lexical ontology. For example, the properties of an entity linked to a term are inherited by subtypes of this entity but not necessarily by other, unrelated types.

8 Conclusions

The idea of the Global WordNet outlined in this paper is motivated by the desire to make the dozens of independently developed wordnets interoperable in a way that benefits crosslinguistic NLP applications, reasoning and inferencing, and linguistic-lexicographic studies. Many open questions remain and may only be answered during the implementation. We are currently embarking on a project that seeks to build wordnets for seven languages from different language families and connect them via a formal ontology. While the coverage will be limited to the domain of ecology and the environment, we expect this effort to be a solid test case on which to build further interoperable lexical resources based on the lexicon-ontology duality.

9 Acknowledgments

Felbaum's work is supported by DTO/IARPA, REFLEX, and the U.S. National Science Foundation.

References

Christiane Fellbaum and Judy Kegl. 1989. Taxonomic Structures And Cross-Category Linking in the Lexicon. In Kenneth de Jong and Yongkyoon No, editor, *Proceedings of the Sixth Eastern States Conference on Linguistics*, pages 93–104, Columbus, Ohio. Ohio State University.

Christiane Fellbaum and Katerina Stathi. 2006. Ididome in der grammatik und im kontext: We bruellt hier die leviten? In K. Proost and E. Winkler, editors, *Von Intentionalitaet zur Bedeutung konventionalisierter Zeichen*. Mouton deGruyter, Berlin.

Christiane Fellbaum, editor. 1998. *WordNet:An Electronic Lexical Database*. The MIT Press.

Christiane Fellbaum. 2007. The ontological loneliness of idioms. In A. Schalley and D. Zaefferer, editors, *Ontolinguistics*. Mouton deGruyter, Berlin.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L Schneider. 2002. Sweetening ontologies with dolce. In *13th International Conference on Knowledge Engineering and Knowledge Management*.

Paul Kay and Charles J. Fillmore. 1999. Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction. *Language*, 75:1–33.

Alfred Kroeber. 1917. *California Kinship Systems*. University of California Coyote Press.

J. LeRoux, K. Moropa, S. Bosch, and C. Fellbaum. 2008. Introducing the african languages wordnet. In J. Csirik, D. Csendes, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Fourth Global WordNet Meeting*, Szeged.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the 2003 International Conference on Information and Knowledge Engineering (IKE 03)*, Las Vegas, Nevada.

Adam Pease and Christiane Fellbaum, 2008. *Formal Ontology as Interlingua: The SUMO and WordNet Linking Project and GlobalWordNet*. Cambridge University Press, Cambridge.

H. Rodriguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, A. Marti, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen, and C. Fellbaum. 2008. Arabic wordnet: Current state and future extensions. In J. Csirik, D. Csendes, C. Fellbaum, and P. Vossen, editors, *Proceedings of the Fourth Global WordNet Meeting*, Szeged.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht.

Anna Wierzbicka. 1985. *Lexicography and conceptual analysis*. Karoma Publishers, Inc.

TagParser: well on the way to ISO-TC37 conformance

Gil FRANCOPOULO
TAGMATICA
126 rue de Picpus
75012 PARIS FRANCE
gil.francopoulo@wanadoo.fr

Abstract

We present rapidly the family of standards that are currently under development within ISO-TC37. Then as an example of application of these ISO specifications for French, a concrete industrial parser is described: TagParser.

1 Introduction

The production, processing, use and re-use of linguistic data form a timely and costly part of the daily work of NLP industry and research teams.

Officially recognized specifications are needed. The ISO-TC37 work started from the GENELEX (Antoni-Lay), EAGLES, MILE (Bertagna) and TEI reports and we think that the family of specifications currently developed within ISO-TC37 is a good help as a common mechanism for fostering interchange of language resources and linguistic processing tools.

As a matter of fact, the title of this paper is not "TagParser: ISO-TC37 standard conformance" because most of the specifications developed within ISO-TC37 are not ISO standards yet. Let us recall that to be called "standard", a normative ISO document must have the status designated as "International Standard" with regards to the internal ISO process. An ISO document starts as a "Working Draft", then it becomes "Committee Draft", "Draft for International Standard", "Final Draft for International Standard" and finally "International Standard"[1]. At each step of the process, the document is balloted by the National Member Bodies (i.e. the organizations for normalization in each country), quite often technical comments are expressed and a new version of the document is produced taking into account the comments for the next round. The process is quite long and burdensome but the aim is to let the time in order to fully study the subject and reach a technical consensus.

In this paper, we describe a parsing scheme for French that has been developed during a period of twenty years (on a part time basis) and that has been recently modified according to the emerging TC37 specifications (Francopoulo 1988, 2005). A version for English has recently been developed and set up with the same strategy.

2 Interoperability requirements

Data associated with language resources are collected and stored in a wide variety of formats. These differences in approach inevitably lead to variations that prevent interchange and re-use of data.
Data are coded according to the following different levels:

- a physical level such as RDF schema (RDFS) or basic XML tags;

[1] see: www.iso.org + processes and procedures

- a basic constant level for values like character set representation (e.g. Unicode), country codes (e.g. ISO-3166), language coding (e.g. ISO-639-3), script coding (e.g. ISO-15924) that are already defined and stabilized outside ISO-TC37.
- a data category level for linguistic constants like /feminine/[2] and attribute values like /grammatical gender/;
- a structural level in which the organization of the data categories are determined in terms of classes and relations among classes. For instance, an entry in a lexicon will be coded according to a class specification holding an attribute called /part of speech/ and will be linked to one or several senses, that are themselves coded according to another class specification.
- a linguistic level such as annotation guidelines that specify what is the rationale that gives one or two NPs in the fragment: "Le député Robert Dupont …". Obviously, this is specific to a given language.
- the quality of the linguistic descriptions in terms of accuracy, coverage and depth with regards with the given concrete language.

Within ISO-TC37, a collective work is in progress in order to elaborate a family of specifications in order to improve current interoperability among language resources. But this work does not deal with all the levels mentioned hereby. The physical level is subject to debates and not fixed: so, one or several schemas are given in the informative annexes[3] of the different ISO standards. The basic constant level is already stabilized and widespread, thus these standards are just referred and, on purpose, no attempt is made to deviate or to redefine these values. Linguistic level and quality are considered as out of scope and thus, not addressed.
Data category and structure level are, on the contrary, the main focus of ISO-TC37 work by means of two kinds of normative objects:

- a data category registry (DCR) (Ide 2004)[4].
- a family of four structural specifications for lexicons and annotations.

TagParser has been re-engineered according to these standards, and seems to be one of the first parsing scheme for French that has this property.
One should note that to be interoperable, a parsing scheme does not need to directly implement all these standards. The TC37 specifications are designed as pivot formats implementing an abstract data model for lexicons and annotations. Thus, only mapping is needed. But why not using the ISO standards as direct guidelines? First, it is more simple to offer a direct interoperability instead of using mapping and secondly, the specification being rather generic, using them as a foundation offers a good guarantee for future evolution.

3 Data category level: data categories recorded in a registry

Data categories include both attribute such as /grammatical gender/ as well as a set of associated atomic values, such as /feminine/. In both cases, the abstraction behind an attribute or value is distinguished from its realization as some string of characters. To serve the needs of the widest possible user community, the DCR is developed with an eye toward multi-lingualism with the following criteria for each entry:

- an entry identifier;
- textual reference definitions which are expressed in various languages;
- names, possibly with synonyms, which are declared in various languages;
- possibly a shallow ontology is organized in order to link generic-specific values like /common noun/ vs. /noun/;
- possibly for some data categories dedicated to attributes, a range of permitted values.

An important property to mention is that data categories for lexicons and annotations are not separated. Of course, some values are specific to annotation, for instance, /punctuation/ that is mandatory

[2] Following ISO-12620 revision, data category identifiers are expressed between slashes.
[3] One should note that an ISO document comprises two sorts of sections: normative parts and informative parts, the latter being there only to help the understanding and usage of the normative parts.
[4] Data Category Registry: http://syntax.inist.fr

for annotation and is not for some lexicons. A second aspect deals with interoperability: with a set for lexicons and a set for annotation, the danger was too high to face a balkanization and thus to have incompatible sets.
Another point to mention is that the number of values is rather high, currently 600. Thus, the TC37/SC4 management decided to split the work into four sub-tasks on a linguistic basis and not on an object target basis.
So, four ISO profiles (each one corresponding to a sub-task) have been created:

- meta-data[5]
- morpho-syntax
- syntax
- semantics

And all these values are to be shared by lexicons and annotations. Currently (Fall 2007), a set of 600 data categories has been recorded in the ISO data category registry based on the work of:

- EAGLES for West-European languages;
- MULTEXT-East for East-European languages;
- Sfax University for Semitic languages;
- IMDI for meta data values;
- joint ISO-LIRICS-SIGSEM work and TimeML for semantic values;
- different TC37/SC4 works on lexicons and annotations for a small set of values.

One should note that two additional works are currently conducted (in Asia within the NEDO project for Asian languages and in South Africa for African languages) but the values are not yet entered in the database.

4 Structural level: a family of four specifications

So, the data categories provide a good foundation for interoperability between TC37 specifications and external formats, but, of course also among TC37 specifications.

The objects that we deal with are lexicons and annotations, the latest being either the result of a text hand-coding or the output produced automatically by a program. No distinction is made between the two types of annotations.
Four structural specifications are concerned:

- Lexical Markup Framework (LMF) that is the ISO specification for NLP lexicons (Francopoulo 2006)[6]. Individual instantiations of LMF may include monolingual, bilingual or multilingual resources. The same specifications are to be used for both small and large lexicons. The covered languages are not restricted to European languages but cover all natural languages. The descriptions range from morphology, syntax, and semantics to translation. An important part of LMF is dedicated to multilingual notations in order to both link senses of different languages, but also to control translation through a general ontology such as SUMO.
- Morpho-syntactic Annotation Framework (MAF) that first allows to segment a text into tokens and words, and secondly to mark these segments with values like /part of speech/ or /feminine/ (Clément 2005).
- Syntactic Annotation Framework (SynAF) that first rules how to delimit and mark syntactic phrases and sentences, and secondly rules how to describe relations between these phrases (Declerck 2006). SynAF annotations are built on top of MAF annotations. The sentence defines the boundaries of the fragments of textual documents to which SynAF applies.
- Semantic Annotation Framework (SemAF) that specifies how to add semantic marks to a text[7]. SemAF relies on MAF and SynAF.

5 One should note that the term "meta-data" for this profile is a bit misleading because, in fact, all data categories may be used as meta-data. This profile covers management marks like /creation date/ and /author/.

6 see: www.lexicalmarkupframework.org

7 Contrary to MAF and SynAF, SemAF is a multipart specification and is not very well developed. Among the various parts that are scheduled, only the part one, that deals with time and events is active.

5 TagParser

5.1 Architecture

Two important points need to be mentioned in order to understand the parsing pipeline.
First, the parser itself is not a stand-alone program. The parser is just one of the components of a full parsing scheme that comprises modules like format guesser, format reader, language guesser, segmentation module, morphological analyzer, named entity recognizer, unknown word guesser and spelling corrector. All these modules implement a 'stand off' notation strategy as described in MAF ; **that is each module computes an annotation that is added layer by layer.** That means that the original textual content can be referred by means of pointers in all layers.
Secondly, TagParser proceeds in two main steps:

- a hybrid (symbolic and statistical) chunker that is corpus-based;
- a constraint solving module that is rule-based.

Oddly enough, there is no part of speech tagger. Using a tagger as a first pass for a parser is not very well suited for French. We know now that, since the GRACE campaign where 21 programs were compared with the objectives of tagging various sorts of texts. In this campaign, the winner was not a tagger but a robust chunker (Adda 1999, Vergne 2005). This can be explained by a certain number of reasons. One is that taggers usually operate on a window of two, three or four words, but in French, we have frequently various phenomena whose scope is broader. Another aspect is the significance of frozen multi-word expressions (MWE) that do not respect regular grammatical behavior, and so do not conform to a simple statistical model. The main problem for taggers in French is that they give too many wrong results. Ten years ago, when parsers had F-scores[8] in the range of 50 - 60%, this error rate was not a serious problem, but now, where parsers are more in the range of 70 - 90% or higher, this error rate is proportionally much more important. In the community, a familiar proverb is : "using a tagger for a parser is like starting to work by shooting oneself in the foot".

This does not mean that statistical methods cannot be used for French. This just means that the notions of chunks and MWEs must be taken into account.

5.2 Lexicon

To this regard, an essential element is the lexicon. TagParser is associated with an LMF conforming lexicon comprising 600 K inflected forms obtained from 100 K simple lemmas and 30 K MWEs, these latest ones covering most frequent idiosyncrasies. The syntactic descriptions come mainly from DicoValence (Van den Eynde 2003). Here is an extract of the lexicon:

```
<LexicalResource dtdVersion="14">
   <GlobalInformation
      <feat att="languageCoding" val="ISO 639-3"/>
   </GlobalInformation>
   <Lexicon>
      <feat att="language" val="fra"/>
      <LexicalEntry paradigmPatterns="AsPassif">
        <feat att="partOfSpeech" val="adjective"/>
        <Lemma>
           <feat att="writtenForm" val="actif"/>
        </Lemma>
        <Sense id="S1">
           <feat att="definition"
                 val="Qui agit ou implique une activité"/>
           <SenseRelation targets="S3">
                 <feat att="label" val="antonym"/>
           </SenseRelation>
        </Sense>
        <Sense id="S2">
           <feat att="definition"
                 val="Propre à exprimer que le sujet est considéré comme agissant"/>
           <feat att="domain"    val="grammaire"/>
        </Sense>
        </LexicalEntry>
     <LexicalEntry paradigmPatterns="AsPassif">
        <feat att="partOfSpeech" val="adjective"/>
        <Lemma>
           <feat att="writtenForm" val="inactif"/>
        </Lemma>
        <Sense id="S3">
           <feat att="definition"
                 val="Qui n'a pas d'activité"/>
        </Sense>
        <Sense id="S4">
           <feat att="definition"
                 val="Qui n'a pas d'activité régulière, sans être chômeur"/>
           <feat att="domain"    val="juridique"/>
        </Sense>
        </LexicalEntry>
   …
</LexicalResource>
```

[8] The harmonic mean of precision (P) and recall (R): i.e. F-score = (2 * P * R) / (P + R)

The element "AsPassif" is a shared paradigm pattern defined elsewhere in the lexicon in order to describe that the lemma "actif" gives the inflected forms "actif", "actifs", "active" and "actives" for the four combinations of number and gender.

5.3 Coverage

Another aspect for a parser for a given language is to determine the corpus for this language. In France, we do not have any reference corpus like the British National Corpus or the American National Corpus for English. So, an attempt has been made to approximate a "balanced" corpus. The corpus is made of 82 M words: 65% of the texts comes from various news sources (belonging to general, sport and economic genres), 30% comes from parliamentary minutes (coming from both EC and French institutions), 4% comes from literary sources and 1% comes from emails and oral transcriptions of street dialogues. The proportions are rather open to criticism but this is all what we could collect. For us, this corpus is what we call "the French language". Let us add that this corpus is a raw corpus: there is no annotation of any kind.

5.4 Chunker development process

The main part of the parser is the chunker. This module has the difficult task of splitting the sentence into chunks, labeling these chunks and tagging part-of-speech ambiguities. The chunker produces only one solution.

The task of developing a rule-based chunker is a rather difficult one. The maintenance of a set of rules turns rapidly into a nightmare. We decided a long time ago to adopt a more stable strategy that is to induce a chunker from unordered examples. The question was how to select examples? The task of hand-coding annotation is a rather time consuming one. Thus, the annotation of randomly selected examples with the objectives of having a broad coverage is out of reach. The best strategy is dynamic annotation selection.

In this process, the parser is incrementally improved through a series of small steps.

Dynamic annotation selection algorithm: A tiny hand-coded corpus is used to serve as a bootstrap to build an initial parser by means of a machine learning algorithm. The parser is then applied to the raw corpus. Parsing failures are collected and the most simple failures are ranked. Similar situations are pruned. And the related sentences are then hand-coded and added to the hand-coded corpus. The system is then ready for another step (Francopoulo 2003).

In fact, it is a little bit more complex than that because the learning process has its own inner loop. Each time a new parser version is induced, an automatic check is made to ensure that the new version is at least better than the last one. This check is done by applying and evaluating the induced parser to the hand-coded corpus. Most of the time, the quality is better but if it is not the case, the situation is intellectually studied that may lead to lexicon accuracy improvement, tagset refinement or annotation guidelines modifications (that may lead themselves sometimes to backwards corpus updates). So, the process is globally incremental but some problematic situations may conduct to move temporarily one step behind, before going forward and further[9]. The algorithm pertains to the family of data-oriented parsing (DOP) when applied to chunks, on the contrary to DOP schemes applied to trees (Bod 2003).

The hand-coded corpus contains currently 90 K words and allows the parser to cover 96% of the raw corpus. Obviously, the hand-coded subset is not a corpus that is representative of the French language from a numerical point of view because the proportions are clearly biased. This corpus is more to be considered as a collection of difficulties.

The machine learning algorithm does not operate directly on data categories coming from the MAF result but on tagsets that are defined as combination of ISO data categories. Currently, the number of tagsets is 205. Most specific French grammatical words have their own tagset because these words exhibit rather specific combinations within phrases like NP or VP.

5.5 Constraint solving module

The machine learning mechanism is applied to chunks and works fine but, this strategy is difficult to apply to computation of syntactic relations because the annotated corpus is too small. A set of

[9] These notions are usually called local optimum vs. global optimum in the context of meta-heuristics like simulated annealing.

constraint rules have been hand-coded instead. The constraints combine grammars using strictly local rules with syntactic information obtained from DicoValence. The rules are organized into 14 packages, each of them implementing one of the relations as expressed in the PEAS guidelines for French[10]. A constraint solving module is applied during the parsing process.

5.6 Format of the result

Following MAF and SynAF, the result comprises six levels: Token, Word Form, Group, Relation, Sentence and Document.
A **token** is character string defined by an algorithm dedicated to segmentation.
A **word form** is defined as the result of :

- a named entity recognition,
- a look-up in the lexicon,
- or an unknown word guessing.

A **group** is a contiguous sequence of word forms. Aside from a limited number of specific situations[11], a group is the result of the chunking process. A group is non-recursive. The labels of the groups are taken from the DCR and the list is as follows:

- verbNucleus
- nounPhrase
- prepositionPhrase
- adjectivePhrase
- adverbPhrase
- prepositionVerbPhrase

A **relation** is a link between word forms and/or groups. The labels of the relations are also taken from the DCR and the list is as follows:

- subject
- auxiliary
- directObject
- verbComplement
- verbModifier
- complementizer
- attribute
- nounModifier
- adjectiveModifier
- adverbModifier
- prepositionModifier
- coordination
- apposition
- juxtaposition

A **sentence** is defined as the contiguous sequence of word forms linked by the transitive closure among relations.
A **document** is defined as the whole set of sentences in a file.

5.7 Implementation and speed

The code is written in Java for the development tools as well as for the parsing pipeline. Like most modern Java industrial codes, the multi-core and multi-processor features of recent computers are exploited when available. More precisely, the number of cores and processors is consulted at start-up time and accordingly, a certain number of parsing processes are run in parallel, the lexicon being loaded only once.
The learning phase together with the self-check phase takes 10 minutes. The whole parsing pipeline has a speed rate of 600 K words per hour on a server class machine (mono-Xeon quad-core). This speed is usually considered as acceptable for industrial purposes.

5.8 Evaluation

TagParser competes in the evaluation campaigns of the ANR-Passage project (see acknowledgements). The first evaluation will be conducted in December 2007 on a 'black box' basis.

The objectives of this project are also to build a 200 M words annotated corpus for French based on the combination of ten parser results. This project is a French National campaign that gathers most of the known parsers for French. The corpus will respect ISO-SynAF specifications, but it is still a bit too early to present any concrete result.

[10] see: www.limsi.fr/Recherche/CORVAL/easy
[11] In French, a chunk beginning with "de" cannot be distinguished as being NP ("Robert mange de la salade") compared to PP ("Robert arrive de la cuisine") from a syntactic computation based only on word constituency.

6 Conclusion

In this paper, we presented rapidly the family of standards that are currently under development within ISO by a great number of people coming from different countries.
Then, TagParser was described as an example of ISO specifications application.

Acknowledgements

This work was supported in part by the EU (eContent project 22236 LIRICS[12]) and in part by the French ANR-Passage project (Action ANR-06 MDCA-013[13]).

References

Adda G., Mariani J., Paroubek P., Rajman M. 1999 L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. Langues vol-2.

Antoni-Lay M-H., Francopoulo G., Zaysser L. 1994 A generic model for reusable lexicons: the GENELEX project, Literary and Linguistic Computing 9(1): 47-54

Bod R. 2003 Extracting stochastic grammars from treebanks, in Treebanks: building and using parsed corpora, Abeillé ed, Kluwer

Bertagna F., Lenci A., Monachini M., Calzolari N. 2004 Content interoperability of lexical resources, open issues and MILE perspectives. LREC Lisbon

Clément L., de la Clergerie E. 2005 MAF: a morpho-syntactic annotation framework. Language & Technology Conference Poznan

Declerck T. 2006 SynAF: towards a standard for syntactic annotation. LREC Genoa

Francopoulo G. 1988 A parser for French with induction of grammar rules, PhD dissertation, Paris-6 University

Francopoulo G. 2003 TagChunker: mécanisme de construction et évaluation. TALN Batz sur mer

Francopoulo G. 2005 TagParser et Technolangue-Easy TALN Dourdan

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF) LREC Genoa

Ide N., Romary L. 2004 A registry of standard data categories for linguistic annotation LREC Lisbon

Van Den Eynde K., Mertens P. 2003 La valence : l'approche pronominale, application au lexique verbal. Journal of French Language Studies 13, 63-104

Vergne J., Houden F. 2005 L'analyseur syntaxique Vergne-98 présenté aux actions d'évaluation GRACE et EASy. TALN Dourdan

[12] see: http://lirics.loria.fr
[13] see: http://atoll.inria.fr/passage

TIMEML: An ontological mapping onto UIMA Type Systems

Del Gratta Riccardo, Caselli Tommaso, Nilda Ruimy and Nicoletta Calzolari
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
via Moruzzi 1 -56124- Pisa, Italy
{riccardo.delgratta,tommaso.caselli,nilda.ruimy,nicoletta.calzolari}@ilc.cnr.it

Abstract

We present *TeR*, an UIMA Type System (Ferrucci and Lally, 2004) for event recognition, for temporal annotation in an Italian corpus[1]

We map each TIMEML category (Pustejovsky et al., 2006) to one or more semantic types as they have been defined in the SIMPLE-CLIPS ontology (Ruimy et al., 2003). This mapping presents some advantages, such as the orthogonal inheritance that an event can acquire when derived from the ontology and a clearer definition of semantic roles when carried out by events.

The mapping is implemented by means of a FINITE STATE AUTOMATON which uses semantic information collected from the SIMPLE-CLIPS ontology to analyze natural language texts.

1 Introduction

Temporal information has become one of the key points in Computational Linguistics and Semantics Research fields. A lot of studies pointed out that the understanding and treatment of "time" in texts play a crucial role both theoretically, (e.g: for "bridging anaphora" resolutions and phrase structures), and from an applicative perspective, (e.g: Open Domain Question-Answering, Information Retrieval, Information Extraction, Semantic Web, etc.) (Saurí et al., 2005).

In natural language texts, events are strongly anchored in time. It is by using time and temporal relations between events that, as human, we can reason on changes of certain situations (Hobbs and Pustejovsky, 2003). Temporal relations among different entities represent the core elements to describe the temporal ordening of a situation. Moreover, temporal ordening understanding is also responsible for reasoning.

By annotating events we can draw a kind of map over the text itself which makes easier the access to the information through temporal context rather than keywords.

In this article, we present an "UIMA-based" resource interoperability achieved by integrating the PAROLE-SIMPLE-CLIPS lexical resource (Ruimy, 2006) with both TIMEML annotation guidelines and rule-based heuristics. One of the main advantages of using the UIMA platform is that it allows the "embedding" of already existing resources in the framework, rather than the implementation of new ones and the possibility to easily integrate additional tools in an "IDE" such as *Eclipse*.

[1] More information about corpus used, and its characteristics, see (Caselli et al., 2007)

2 Linguistic issues

TIMEML is one of the most complete annotation scheme for temporal annotation and it has been recently proposed as an ISO standard. Its specification languages allows the identification of events and states (e.g: go, try, peace, on board ...), temporal expressions (e.g: December 25th, four years ...), and temporal links among these entities (e.g: after, during ...).

For the purpose of this work, we concentrate only on the event category. TIMEML employs a rather pre-theoretical notion for defining events and states, i.e. as something which happens or occurs, and as situations which hold or obtain to be true. Events are not classified using the particular meaning of the verb which describes them, but through the use of semantic information encoded in the events themselves. Once identified, events are classified in one of the following categories:

- REPORTING
- PERCEPTION
- ASPECTUAL
- I_ACTION
- I_STATE
- STATE
- OCCURRENCE

This classification is useful since it is language independent and relevant for "characterizing the nature of an event as being irrealis, factual, possible, reported, intensional" (Saurí et al., 2005).

Our approach to event recognition and classification, is to link -or better to map- each of the above seven categories to one or more semantic types as they have been defined in the SIMPLE-CLIPS ontology. This mapping provides semantic information because each event is associated with a lexical entry from which it inherits other semantic information, according to the orthogonal inheritance principle (Pustejovsky and Boguraev, 1993).

3 Automatic TIMEML tagging

This section analyzes two of the various attempts performed to semi-automatically recognize TIMEML category (section 3.1) and describes our proposal (section 3.2).

3.1 Background

Attempts to recognize TIMEML categories have been performed by using ontology semantic types (Caselli et al., 2007), or by using TimeBank and machine learning approaches (Boguraev and Ando, 2006).

The strategy followed by Caselli et al. (2007) is to (manually) check whether a word sense[2] denotes an event and which is its ontological semantic type. Heuristics have been used to check the current word sense against its semantic and syntactic properties as defined in the lexical resource for its classification.

The strategy followed by Boguraev and Ando (2006) is a hybrid approach using both a finite state grammar for temporal expressions and a machine learning technique trained on the TimeBank and unannotated corpora. The finite state grammar is embedded in a shallow parser, while the machine learning algorithms implement novelty in learning from unannotated corpora.

3.2 Our proposal

Event recognition has recently reached quite good levels although improvements can still be obtained.

This section explains our approach to (semi-automatically) implement a **T**imeML **E**vent **R**ecognizer, henceforth *TeR*.

Our idea is formally based on the work of Caselli et al. (2007) but the strategy we follow is the opposite.

We use the SIMPLE-CLIPS ontology to answer the question:

when does a word sense denote an event?

In the SIMPLE-CLIPS ontology, each word sense is classified in terms of the semantic type it belongs to. Word senses which belong to the SIMPLE-CLIPS event type system are collected into lists according to the *type* of the event. These lists are then processed to mark-up word senses with the correct TIMEML category which is "suggested", on the one hand, by the semantic type to which the senses belong to, and on the other, by rule-based heuristics. In addition, the use of the PAROLE-SIMPLE-CLIPS

[2] By word sense we mean the sense of the word currently analyzed. Word sense disambiguation is a crucial aspect in (Caselli et al., 2007) work. Also in our proposal this issue has to be addressed.

lexical resource allows TIMEML categories to be enriched with a lot of morphosyntactic and semantic features, directly inherited from the resource.

The chance to uniquely classify events following TIMEML specifications represents an important step toward the implementation of algorithms capable of managing a strong automatic treatment of texts.

The UIMA platform is a focus in our event recognition approach, since it is responsible for providing both the final mapping between the SIMPLE-CLIPS semantic types and the TIMEML categories via rule-based algorithms and heuristics implementation and the common platform on which those resources are integrated.

Our goal is to define a set of UIMA Type Systems useful to identify TIMEML categories in texts.

4 TeR, a TIMEML Event Type System

TeR is a **T**imeML **E**vent **R**ecognizer implemented as an UIMA TYPE SYSTEM used to tag word senses which denote events. The TeR is built upon the PAROLE-SIMPLE-CLIPS lexical resource from which it inherits syntactic and semantic information. This additional information is attached to the word sense and handled as UIMA Type System features of the TeR, which is promoted to be a "complex collector" of several linguistic information as well as the "classifier" of the TIMEML category.

As explained in section 6, TeRs are built by integrating different resources, which are described in the following subsections. Section 4.1 is a brief introduction to the SIMPLE-CLIPS ontology; section 4.2 presents the UIMA Type System hierarchy and features; section 4.3 is a description of the TeRs as "complex collectors" of information and, hence, complete annotators.

4.1 The SIMPLE-CLIPS Ontology

In the PAROLE-SIMPLE-CLIPS lexical database, at the semantic layer of information, lexical units are structured in terms of a semantic type system and are characterized and interconnected by means of a rich set of semantic features and relations. The type system structure consists of 157 language-and domain-independent semantic types designed for the multilingual lexical encoding of concrete and abstract entities, events and properties.

The SIMPLE-CLIPS ontology, as already stated, implements the principle of orthogonal inheritance, whereby multidimensionality is captured by qualia roles[3] (Pustejovsky, 1995) which define the distinctive properties of semantic types and differentiate their internal semantic constituency.

According to the philosophy governing the SIMPLE-CLIPS ontology, a semantic type is the repository of a structured set of semantic information about a lexical unit. In the lexical database, predicative word senses are assigned a predicate-argument structure. Predicative information consists in the description of the argument structure in terms of predicate arity, semantic role and semantic constraints of each argument[4]. It is worth noting that the encoding of restrictions on arguments entails that the lexical resource provides information not only on word senses but also on their semantic context, which is a useful information for event classification according to the TIMEML specifications.

4.1.1 Event tree structure

In SIMPLE-CLIPS ontology, 59 different event types have been defined. These event semantic types are organized in the typical tree structure: we can identify 7 main (root) event types, each one subsuming a certain number of sub-events.
This event tree structure is an "*is-a*" relation with edges and nodes arranged in a tree structure. In the perspective of the FINITE STATE AUTOMATON outlining (see section 5.6), the distance of the node from the upper root[5], as well as its direct ancestor are relevant for rule-based heuristics implementation.

4.2 Defining UIMA Type Systems for TeRs

As a framework for our common platform, we adopted the UIMA Architecture. UIMA provides both the integration of single NLP components, PRIMITIVE ANALYSIS ENGINES, and NLP pipelines deployment facilities, AGGREGATE

[3] In the framework of the SIMPLE-CLIPS ontology model definition, each of the four *qualia* roles has been promoted as top node of a hierarchy of semantic relations which altogether form the *Extended Qualia Structure*.

[4] Constraints are expressed in terms of semantic type, features, or "notions" combining these different expressive means.

[5] Here, the upper root node is the (generic) "Event" .

ANALYSIS ENGINES. The former can be either existing resources or resources built from scratch within the UIMA framework, while the latter represent a NLP workflow defined on the integration of single components via descriptor fence.

The ANNOTATION TYPE, i.e. the TeR upon wich a PRIMITIVE ANALYSIS ENGINE is implemented, that we propose consists of three different kinds of features[6], as reported in table 1:

UIMA TeR feature	**Type of information**
TIMEML Category	Temporal information
Part of Speech Grammatical information Lemma	Morpho-syntactic information
Semantic Type Features Relations	Semantic information

Table 1: TeRs features

Our idea is to set the TeR feature values with information stored in PAROLE-SIMPLE-CLIPS lexical database. Since PAROLE-SIMPLE-CLIPS is a rich resource, each TeR can be enriched with morphological, syntactic and semantic information from the resource, thus becoming a multidimensional object.

We define a super type, *SimpleTeR*, which directly inherits from the *uima.tcas.annotation* type and implements all the features and relations shared among different event types[7].

We define seven different TeRs, one for each TIMEML category. These TeRs inherit from the *SimpleTeR* and record semantic and morphosyntactic information of terms they are annotating as UIMA Type System feature values. The TeRs behave both as standard annotators (including semantic aspects) and as temporal annotators.
Figure 1 shows the hierarchy defined for TeRs. The sentence annotator and its relevance is explained in section 5.7.

[6]For the sake of clarity, we remind that a feature, in UIMA language, is an attribute name-value pair.

[7]At the top level only the formal relation and the link to the Entity class are implemented.

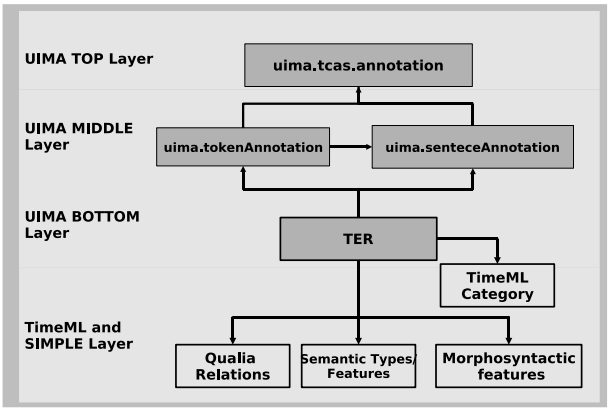


Figure 1: TeRs Hierarchy

4.3 TeR is a "complex collector" of information

An important side effect of retrieving lists of word senses from a lexical resource is that these word senses inherit both semantic properties defined for the semantic type they belong to and morphosyntactic features from the syntactic unit. TeRs have been defined to manage all this complex bundle of information using specific features.
Annotating texts with TeR allows users to perform a multilevel annotation by filling the (UIMA) feature values with semantic and morphosyntactic information which, in turn, enrich the TIMEML information itself.

5 Building UIMA Type Systems: From lexicon to annotators

This section explains some preliminary steps to be performed during the tuning of the lexical resource. Sections 5.1 to 5.3 address the cleaning of the lexical resource, while sections from 5.4 to 5.6 describe the classification of TIMEML categories and the algorithms used to map PAROLE-SIMPLE-CLIPS lexical units onto these categories.

5.1 Cleaning the PAROLE-SIMPLE-CLIPS database: preprocessing

The analysis of the resource shows that a word may have multiple senses and that one sense may have different syntactic behaviors and/or different subcategorization frames. All these ambiguities lead to the word sense disambiguation issue from the semantic perspective and/or to the co-textual analysis from the syntactic perspective.

The PAROLE-SIMPLE-CLIPS database preprocessing step consists in filtering out all event denoting words[8] which are linked to more than one sense and/or have more than one syntactic behaviour. This approach defines an "*a priori*" disambiguation, so that, at the beginning of the annotation process, only unambiguous words are processed.

5.2 Cleaning the PAROLE-SIMPLE-CLIPS database: evaluation of disambiguation

The above defined ambiguities weigh on the whole set of event denoting words in the percentage reported in table 2:

Type of Ambiguity	Percentage
Semantic	37%
Syntactic	21%

Table 2: Ambiguity Distribution

The percentage of syntactic ambiguity reduces to 17% for semantically unambiguous terms.

After preprocessing step, the annotation system is able to *automatically* recognize up to 53% of unambiguous events in texts. By unambiguous event, we mean an event denoted by a word which has one and only one sense and one and only one syntactic behavior.

5.3 Cleaning the PAROLE-SIMPLE-CLIPS database: list of words output format

For unambiguous words we have decided the following output structure:

Morpho-Syntactic information	Semantic information

Table 3: output file structure

In table 3, morphosyntactic information consists of morphological units, subcategorization frame, lemma, inflected forms and grammatical features; semantic information is the set of semantic features and relations which characterize a semantic type.

[8] An event denoting word is a word belonging to the SIMPLE-CLIPS event type system.

This output format is relevant to assign the correct TIMEML category both when only semantic information is needed and when also syntactic criteria are used to select the TIMEML category (see section 5.4). It is worth noting that this format is coherent with the TeR Type System defined in section 4.2.

5.4 Classification of TIMEML categories

Mapping the SIMPLE-CLIPS ontology over the TIMEML categories defines a correlation between the former and the latter. The relation between the ontological types and the TIMEML classes is "one to many". This is due to the fact that the TIMEML categories depend on several criteria consisting either of semantic clues or of a mixture of semantic and syntactic ones.

Heuristics necessary to uniquely identify the final TIMEML category have been implemented to improve the mapping.

All experiments performed suggest that a clear approach must be followed to map the SIMPLE-CLIPS ontology to the TIMEML categories: first semantic information is analyzed and then co-textual information (i.e. verb forms, argument structure or syntactic dependencies) is applied to assign the right TIMEML class. We classify the seven TIMEML categories according to the following rules:

TIMEML Category	Information
REPORTING PERCEPTION ASPECTUAL	Semantic Criteria, including lexical meaning
I_ACTION I_STATE STATE	Semantic, including lexical aspects, plus syntactic criteria
OCCURRENCE	Default exit [9].

Table 4: TIMEML categories and rules

5.5 Coarse-grained mapping

The PAROLE-SIMPLE-CLIPS resource is implemented as a relational database. Each object, implementing semantic properties described in section 4 and other morphosyntactic information, is a specific

[9] In the following we will see that a FS grammar built over SIMPLE-CLIPS ontology always has an "OCCURRENCE" as default exit when no other TIMEML category could be assigned.

table. The structure of the tables and their values define both how to map the TIMEML category onto word senses and how TeRs are built accordingly.

This subsection is dedicated to the outline of the FINITE STATE AUTOMATON (FSA) responsible for the TeRs definition and implementation in a coarse-grained mapping scenario.

Following Caselli et al. (2007), we can implement basic rules which, essentially, rely on the event tree structure. As a starting point of the mapping process we consider only the 7 root-events. Fine-grained investigation over sub-events is performed at a deeper level of analysis, when heuristics are implemented (see section 5.6).

The first "coarse-grained mappings" implemented concern the *Phenomenon* and the *State* semantic types: all word senses belonging to these types and their subtypes are mapped to the TIMEML category of OCCURRENCE and STATE respectively (see table 5):

Event	TimeML	Rule
Phenomenon	OCCURRENCE	Stop at root level
State	STATE	Stop at root level

Table 5: Coarse-Grained Mapping for Phenomenon and State semantic types

The FSA rule can be read as follows:

- `output_step_0=Extract words belonging to Phenomenon (State) from SIMPLE-CLIPS;`
- `output_step_1=Match strings in texts with the output_step_0;`
- `Exit_FSA=Tag output_step_1 with OCCURRENCE (STATE);`

Figure 2 shows the FSA for the coarse-grained mapping.

5.6 Fine-Grained Mapping

Fine-grained mapping combines both linguistic and heuristic-based techniques to better assign the TIMEML category to a given word sense.

In a fine-grained mapping scenario, FSA uses the rules directly resulting from the heuristics, in which

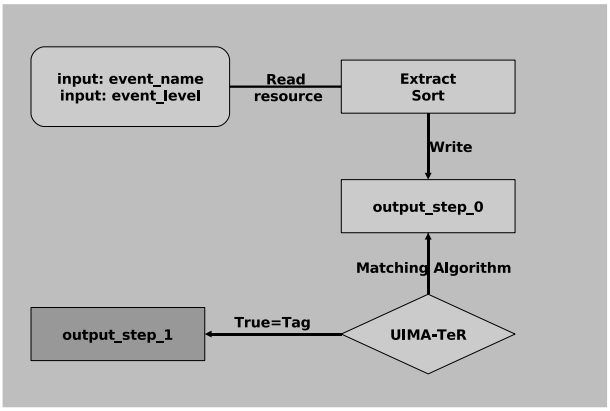


Figure 2: FSA for the coarse-grained mapping.

both semantic and co-textual criteria are used to build guidelines in a hierarchical order of application, whereby the semantic information is dealt with beforehand.
The strategy can be summarized as follows:

- `output_step_0=Create lists of words for each class of events and sub-events;`
- `wip`[10]`_step_0=Add semantic and mophosyntacic information each word in output_step_0;`
- `wip_step_1=Apply heuristics;`
- `output_step_1=Extract from output_step_0 only words uniquely mapped onto TIMEML categories;`

The fine-grained mapping is essentially based upon wip_step_1. In this step, we apply heuristics to each list of words. The internal rules of heuristics allow FSA to take decisions about the TIMEML category a given word sense should belong to. Co-textual analysis and predicative structures in prepositional phrases are key points in fine-grained mapping, since FSA is also driven by the complex phrase structure (see section 5.7).

5.7 Co-Textual analysis and sentence annotator

In order to be as compliant as possible with TIMEML specifications and to manage event variability, we have to consider, also, a portion of text

[10] wip means work in progress.

surrounding the event and the actual realization of the event itself.

Some heuristics check whether an event has another event as its proper argument. To be able to manage this co-textual analysis we implement the "window capability" of the UIMA framework, i.e. the number of tokens which are analyzed in the same annotation process. Window capability allows UIMA to span a large part ot text. The main token of the window is the event denoting word sense and other tokens in the same window are analyzed by an automaton responsible for semantic and morphosyntactic recognition. In this scenario single tokens are relevant not only by themselves, but also for their interactions with other tokens. Since the sentence is a coherent set of tokens, we chose one single sentence as the spanned text in the "window". Within a sentence, the FSA recognizes different syntactic behaviors and semantic restrictions on predicative structure and heuristics can be completely applied.

6 First prototype of UIMA Type System annotation tool

A first prototype for UIMA Type System has been built to manage TIMEML categories which do not need the window size in the text analysis, i.e. they do not need the co-textual analysis to be assigned to the word sense. For example, the *Phenomenon* semantic type belongs to this kind of categories. Input text is passed to UIMA and translated into an object, the CAS (Götz and Suhre, 2004), which contains both physical and metadata of the text itself. The CAS ***initialize*** and ***process*** methods are responsible for cleaning and setting up the PAROLE-SIMPLE-CLIPS resource and for the mapping between the unambiguous word senses and the TIMEML categories respectively.

A second step is the development of the co-textual automaton as independent software. This automaton is called by the UIMA framework every time a window capability is required. The sentence annotator represents the basic step to implement the co-textual automaton. Sentence annotator and TeR are "aggregated" so that, within a single sentence, every token is analyzed by a morphosyntactic parser and the information retrieved is sent back to the UIMA platform to be handled by the ***process*** method of the AGGREGATE ANALYSIS ENGINE responsible to apply heuristics. Figure 3 below shows how the interoperability between the PAROLE-SIMPLE-CLIPS resource and TIMEML specification is implemented. The layout of the figure is based on the Language Grid Service Ontology. The arrows point to objects which are the subjects of the properties: for example in "PAROLE-SIMPLE-CLIPS *usedBy* UIMA FW", the arrow points towards PAROLE-SIMPLE-CLIPS.

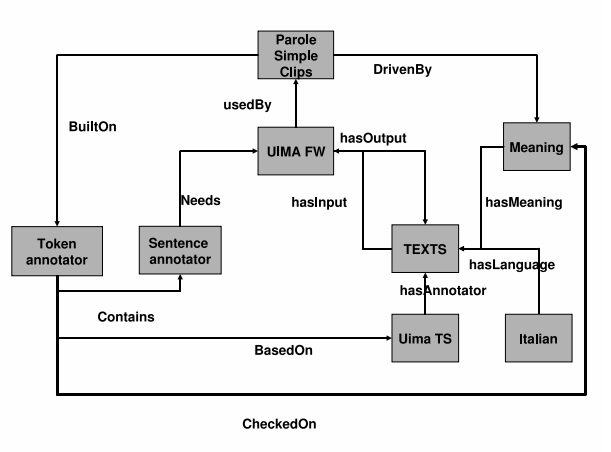


Figure 3: Work Flow

7 Testing the platform

This section briefly describes some test cases that have to be performed after the first prototypes have been deployed.

The corpus to be annotated is the one used by Caselli et al. (2007) for their manual mapping of events. Our results will be checked against their results in terms of numbers of events identified, percentage of agreement on events and overall agreement on classification (K-statistic).

8 Managing the disambiguation

As explained in section 5.2, a word may have different senses.

Since the TIMEML categories are mapped to words via their sense, the disambiguation problem is crucial in TIMEML mapping.

For example, the Italian word "distrutto" in the following sentences:

1. Edificio **distrutto**...[**Destroyed** building ...]

2. Le tue parole mi hanno **distrutto** [Your words **wrecked** me]

has two different senses: in the first sentence the sense is linked to the semantic type *Cause_change_of_state*, while in the second it is linked to the *Cause_experience_event*.

The word "distrutto" has to be mapped onto two different TIMEML categories:

- distrutto$(1) -> $ STATE
- distrutto$(2) -> $ OCCURRENCE

The FSA, in the fine-grained mapping scenario, is able to switch between these two possible senses according to their different syntactic behaviour.

9 Conclusion

We have presented TeR, a TIMEML Event Annotator modeled upon a rich lexical resource like PAROLE-SIMPLE-CLIPS. The "a priori" disambiguation formalized in the resource allows TeR to automatically tag up to 53% of words, since this is the percentage of unambiguous terms in the resource.

Resources interoperability is a focus in this project, and the UIMA Platform is the common framework used to integrate resources. This paper intends to contribute both to a UIMA TYPE SYSTEM standard and to a common framework for resource sharing and interoperability definition. Moreover, an operative workflow in the infrastructure is defined.

Strong links are established with the GrAF (Ide and Suderman, 2007) annotation framework, since the span feature of the GrAF is easily mapped on the begin-end features of the TeR, and with the NICT language grid project (Ishida, 2007), from which our prototype inherits the service ontology environment.

In addition, by adding these Type Systems to the UIMA platform, researchers can use them to add a new annotation layer to already existing corpora e. g. to TreeBanks.

References

Branimir Boguraev and Rie Kubota Ando. 2006. Analysis of timebank as a resource for time-ml parsing. In *Proceeding of LREC-06*, May 2006, Genova.

Tommaso Caselli, Irina Prodanof, Nilda Ruimy, and Nicoletta Calzolari. 2007. Mapping simple and timeml: improving event identification and classification using a semantic lexicon. In *GL2007: Fourth International Workshop on Generative Approaches to the Lexicon*, 10-11 May 2007, Paris.

David Ferrucci and Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.*, 10(3-4):327–348.

Thilo Götz and Oliver Suhre. 2004. Design and implementation of the uima common analysis system. *IBM Systems Journal*, 43(3):476–489.

Jerry Hobbs and James Pustejovsky. 2003. Annotating and reasoning about time and events. *Proceeding of the AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning*, pages 74–82.

Nancy Ide and Keith Suderman. 2007. Graf: A graph-based format for linguistic annotations. In *Linguistic Annotation Workshop, ACL 2007*, Prague.

Toru Ishida. 2007. Nict, language grid & department of social informatics, kyoto university. *http://langrid.nict.go.jp*.

James Pustejovsky and Branimir Boguraev. 1993. Lexical knowledge representation and natural language processing. *Artif. Intell.*, 63(1-2):193–223.

James Pustejovsky, Jessica Littman, Bob Knippen, Robert Gaizauskas Andrea Setzer, and Roser Saurí. 2006. Timeml annotation guidelines. *http://www.timeml.org/site/publications/specs.html*.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Nilda Ruimy, Monica Monachini, Elisabetta Gola, Nicoletta Calzolari, Cristina Del Fiorentino, Marisa Ulivieri, and Sergio Rossi. 2003. A computational semantic lexicon of italian: Simple. *Computational Linguistics in Pisa, Istituto Editoriale e Poligrafico Internazionale*, pages 821–864.

Nilda Ruimy. 2006. A computational multi-layered italian lexicon for hlt applications. In *Proceedings XII EURALEX International Congress, Atti del Congresso Internazionale di Lessicografia*, volume 1, pages 221–227, Torino, 6-9 settembre.

Roser Saurí, Robert Knippen, Marc Verhagen, and James Pustejovsky. 2005. Evita: A robust event recognizer for qa systems. short paper. In *Proceedings of HLT-EMNLP 2005, 700-707*.

Cross-lingual Syntactic Subcategorization Analysis Based on Chinese and English Sentence Pairs

Xiwu Han
School of Computer Science and Technology
Heilongjiang University
hxw@hlju.edu.cn

Tiejun Zhao
School of Computer Science and Technology
Harbin Institute of Technology
tjzhao@mtlab.hit.edu.cn

Conghui Zhu
School of Computer Science and Technology
Harbin Institute of Technology
chzhu@mtlab.hit.edu.cn

Abstract

This paper describes an experiment and the results of cross-lingual syntactic subcategorization analysis based on Chinese and English sentence pairs. First, sentence pairs with possible parallel predicates are extracted. Then, 654 bilingual basic types of subcategorization frames are acquired by means of a heuristic acquisition method. Analysis on the results shows that the acquired bilingual subcategorization frames are statistically and syntactically compatible.

1 Introduction

Researches on subcategorization acquisition for a single language have met with considerable achievements since (Brent 1993)'s pioneering work, e.g. those of English (Korhonen 2001), German (Schulte im Walde 2002), Spanish (Chrupala 2003), Czech (Sarkar and Zeman 2000), Portuguese (Gamallo et. al 2002), and Chinese (Han 2005). However, relevant cross-lingual phenomena are only dealt with theoretically in a few remarks lying sparsely in linguistic books for translation (Baker 1992) or second language acquisition (Ellis 1997).

According to early definition of (Chomsky 1965), subcategorization is the process that further classifies a syntactic category into its subsets, and the function of strict subcategorization features is to appoint a set of constraints that dominate the selection of verbs and other arguments in deep structure. Although subcategorization frames (SCF) are generally regarded as functional distributions integrated with both syntactic and semantic information, both concrete definitions and formats for SCF vary greatly from one language to another. This may be the most severe obstacle in cross-lingual subcategorization research.

Subcategories, nevertheless, exist universally in almost all natural languages for almost all linguistic categories. Thus it remains an interesting and challenging question how much for subcategorization theories is common, comparable or compatible cross-lingually. And the answer to this question will surely benefit cross-lingual information processing tasks such as machine translation, which is also our motivation for this study.

In this paper, we will describe an experiment and the results of cross-lingual syntactic subcategorization analysis based on Chinese and English sentence pairs.

Section 2 defines our syntactic subcategorization frames for English and Chinese. Section 3 introduces our corpus and proposes a combined method to recognize possible parallel predicates. Section 4 describes the process and techniques of cross-lingual subcategorization acquisition. Section 5 analyzes the results from a linguistic viewpoint. And the conclusion is given in Section 6 with some suggestion for further work.

2 Syntactic Subcategorization

The accounts of subcategorization phenomena can be either syntactic or semantic or both.

According to the summary in (Korhonen 2001), the full description of verb subcategorization generally consists of seven kinds of linguistic knowledge: a) the number and type of arguments that a particular predicate requires, b) predicate sense in question, c) semantic representation of the particular predicate-argument structure, d) mapping between the syntactic and semantic levels of representation, e) semantic selectional restrictions or preferences on arguments, f) control of understood arguments in predicative complements, and g) diathesis alternations.

Among these, a) and b) are essential information, c) can be regarded as reclassification of the first two in terms of syntactic meaning, d) refines the first three on linking level between syntax

This study is jointly supported by CNNSF (60773069), CPSF (20060400246), and HPFA(LBH-Z06217).

and semantics, e) and f) supplement a) semantically and pragmatically, and the last constitutes a sort of equation relationship on the whole set of SCFs.

In the practices of acquisition, the actual SCF definitions or formats are often various accordingly with the concerned languages. This variety comes firstly from the representations for argument types. Table 1 gives an example for different argument realizations in typical researches for English, German, Spanish and Chinese. According to some linguistic prescriptive routines, NP stands for noun phrase, PP for prepositional phrase, ADJP or JP for adjective phrase, S or SS for clause, SUBJ for subject of the sentence, OBJ for object of the sentence, COMP for complementary phrase, etc[1]. We can see that, for English, Spanish and Chinese, most of the types are syntactic constituents except the functional labels of SUBJ, OBJ and COMP, whereas some of the German SCF arguments are defined by event cases.

Table 1. Example for Argument Types

Language	Argument Types
English (Korhonen 2001)	ADJP, ADVP, NP, S, INF, ING, WH, PP, POSSING, TOBE, SUBJ, OBJ, MP
German (S.S.im Walde 2002)	nominative, dative, accusative, reflexive pronouns, prepositional phrases, expletive *es*, non-finite clauses, finite clauses, copula constructions
Spanish (G. Chrupala 2003)	NP, PP, COMP, ADJP, Pronouns, INF
Chinese (Han 2005)	NP, VP, PP, MP, JP, SS, QP, BP, TP, BAP, BIP

Table 2. Example for Argument Organizations

English SCF (Korhonen 2001)	26. NP-ADJP-PRED / 46 (SUBCAT OC_AP, SUBTYPE RAIS) / XTAG:Tnx0Vs1 she_PPHS1 considered_VVD him_PPHO1 foolish_JJ (VSUBCAT NP_AP)
Chinese SCF (Han 2005)	No.:67/Abr.:NvJN/Cnt.:327 SCF: NP V JP NP |{0,1,0,1,0,1} Prb.:0.006990854|0.007841915 Example1(Verb:装饰): 服务员/nc 们/k 装饰/vg 好/a 了/LE 房间/ng 。/wj Example2(Verb:准备): 同学/nc 们/k 就/d 都/d 准备/vg 好/a 了/LE 课本/ng 和/CNJ 练习本/ng 。/wj

[1] For detailed descriptions and explanations please refer to the original papers.

Secondly, the organizations of arguments also cause great differences. The example in Table 2 shows (in the shadowed parts) that the English SCF No. 26 is made up of both COMLEX and ANLT labels and a semantic subtype 'RAIS', while the Chinese SCF No. 67 contains only syntactic arguments. Other subtypes in English SCF are 'EQUI', 'PVERB', 'DMOVT', 'EXTRAP', 'NONE', etc.

It seems that there is no easy solution to the problem how much subcategorization should constitute syntactically or semantically. As for computational linguistics, we think that a task-oriented method would be a better though expedient answer.

The tasks of subcategorization acquisition and the relevant information application all call for the analysis of concerned SCFs, which actually is a procedure of cognition that induces nature from representations. Therefore, SCF had better to formalize the external information of syntactic functions, i.e. the observable syntactic features. And with facilities of some practical NLP tools, such as taggers and parsers, syntactic SCFs would be more easily acquired than semantic ones. In turn, better acquisition results and more compatible formats would make it more practical for cross-lingual analysis.

Thus, we adopted syntactic subcategorization frames for our special task.

Our 138 Chinese basic SCF types come from (Han 2005) since they are purely syntactic. Appendix A shows the 82 English basic syntactic SCFs, 77 of which were manually regrouped from (Korhonen 2001)'s types (represented by the numbers behind the double slashes), and 5 (labeled with '??') were formed according to real corpus with no counterparts in (Korhonen 2001). The argument types for our English syntactic subcategorization are listed in Table 3, where AS stands for 'as', IT for 'it', RP for particles, and so on.

Table 3. Syntactic Argument Types for English

AP	DP	SS
AS-AP	IT	TO-VP
AS-IF-SS	NP	VP
AS-NP	PASS-VP	VPING
AS-PASS-VP	PP	WH-SS
AS-VPING	RP	WH-TO-VP

3 Recognizing Parallel Predicates

Most existing subcategorization researches are now focused on predicative verbs, and thus it would be more maneuverable and comprehensi-

ble for our cross-lingual analysis to begin with bilingual sentence pairs with parallel predicates. Hence there comes up the task to select those sentences with parallel predicative verbs and to recognize the two parallel predicates within the sentence pair.

3.1 Our Corpus

The corpus used for our experiment consists of 650,000 bilingual sentence pairs of English and Chinese, which were gathered either from public and free Internet resources or our translation works. The sentences are either translated from Chinese to English or vice versa.

To facilitate the predicate recognition and SCF acquisition, before the experiment we parsed the corpus, with English sentences by (Collins 1999)'s head-driven parser and Chinese sentences by the head-driven parser of MI&TLAB in Harbin Institute of Technology (Cao 2006).

We define the parallel predicates as those translatable or translated predicative verbs in the sentence pairs. According to our rough estimation, the English parser has recalled about 96.5% predicative verbs while the Chinese parser only recalled nearly 75%. Therefore, some heuristic techniques should be resorted on for the recognition of bilingual parallel predicates.

3.2 Method of Bilingual Verb Dictionary

The seed version of our bilingual verb dictionary is made up of 19,112 entries drawn from the bilingual dictionary for the Chinese-English machine translation system of CEMT2K developed by MI&TLAB. We extended the seed with synonyms from English WordNet v. 1.2 and Chinese Extended Tongyicicilin v. 1.0. The entry in our final dictionary in turn is organized as bilingual verbal synonym classes, and there are altogether 3,611 entries including 67,836 Chinese and English verbs.

The algorithm for recognizing parallel predicates is described as follows.

For each sentence pair

- *Specify the English predicate V_e;*
- *Form the Chinese predicate candidate set S_c with all potential words, such as verbs or adjectives;*
- *For each candidate V_c in S_c*
  - *Accept the pair $< V_e, V_c >$ as parallel predicates if they appear in one entry of the bilingual dictionary.*

Manual analysis on 15,000 sentence pairs shows that for this method the precision ratio is 86.5% and the recall ratio is only 27.85%. The low recall is obviously due to the limitation of our bilingual dictionary.

3.3 Method of Syntactic Compatibility

Although the diversity of grammatical categories tends to be great, some common word classes, such as nouns, pronouns, verbs, adjectives, etc, mainly constitute the vocabularies of most natural languages. And our observation on English and Chinese parallel corpus also shows that the more literal the translation is, the more counterpart grammatical categories the pair of sentences may share.

We thus define the cross-lingual syntactic compatibility as ***D***.

$$D = \sum_{i=1}^{n} \lambda_i \frac{Min(|GE_i|,|GC_i|)+1}{Max(|GE_i|,|GC_i|)+1}$$

$\boldsymbol{GE_i}$ is an English grammatical category, $|\boldsymbol{GE_i}|$ is the number it occurs in the English sentence, and $\boldsymbol{GC_i}$ is the Chinese counterpart. $\boldsymbol{n}$ is the number of common grammatical categories that make differences in the special task of recognizing parallel predicates. $\boldsymbol{\lambda_i}$ is the weight for the concerned category, which is trained by a simple gradient descent algorithm on a sample of 10,000 manually analysed sentence pairs.

For this recognition task, we employed a maximum likelihood estimation filtering method with a threshold of 0.79 on the syntactic compatibility. Candidate predicate pairs would be accepted as granted if the syntactic compatibility between the related English and Chinese sentences surpasses the threshold.

Evaluation on a sample of 5,000 sentence pairs shows a precision ratio of 79.4% and a recall ratio of 53.94%.

3.4 Combination of the Two Methods

We combined the two methods mentioned above to obtain a larger useful corpus. It is very interesting that the intersection between the recognizing results of the two methods accounts only a very small part, which is about 15.6% of all the recalled sentence pairs. The combined recognition results achieved a precision ratio of 81.3% and a recall ratio of 71.04%.

Further analysis on the sampled corpus shows that the unrecalled sentence pairs with parallel predicative verbs are mainly due to bad segmentation of Chinese verbs or bad parsing results of

the English sentences. Whereas, those sentence pairs with no parallel predicative verbs are usually free transcriptions or bad translations.

4 Cross-lingual Subcategorization Acquisition

The typical framework for SCF acquisition tends to consist of four components, i.e. the preprocessor (often involving a statistical parser), the pattern extractor (to analyze the argument types for related syntactic categories and head lemmas), the SCF hypothesis generator (mapping the patterns with basic SCF types), and the statistic filter (evaluating sets of SCFs gathered for a predicate).

As mentioned in the previous section, we used (Collins 1999)'s parser and (Cao 2006)'s parser respectively for pre-processing English and Chinese sentences. And the following subsections will describe the rest of the components of our cross-lingual acquisition framework and the experiment results.

4.1 Argument Pattern Extraction

For the Chinese sentences, we used the rule-based analyzer of (Han 2005) to extract possible patterns of syntactic argument types governed by the predicate verb, and the argument token precision is 86.5%.

As for the English sentences, we manually analyzed parsing results of the 180 typical sentences given in the Appendix A of (Korhonen 2001), and drawn 66 rules for the pattern extracting task. The argument token precision of this rule-based extractor was estimated to be 92.6% on an open set of 1,000 English sentences from our corpus.

4.2 Heuristic SCF Hypothesis Generation

The function of an SCF hypothesis generator is to classify the previously extracted argument patterns into basic SCF types and reject a few patterns as unclassifiable, and the generator is usually designed according to linguistic knowledge about the predicative verb.

Employment of two independent monolingual generators obviously is the simplest way to generate bilingual SCF hypotheses. The hypothesis token precision of (Han 2005)'s generator was estimated to be 79% on our Chinese corpus, while the English hypothesis generator we designed came up with a token precision of 84.44%. In turn, the potential bilingual hypothesis token precision would at most reach 66.71%, and the actual rate was only 49.2%.

To obtain a higher bilingual hypothesis token precision, we adopted a heuristic hypothesis generation method, which is based on the assumption about ontological arguments.

We define an ontological argument as one obligatory argument headed by a contentive word. For example, in the following two sentences labeled with syntactic arguments, NP and VP are ontological arguments, while PP is not.

a. NP[he] V[saw] NP[a little boy] PP[in the park] PP[with a telescope] .

b. NP[他] PP[用望远镜] V[看见] PP[公园里] VP[有] NP[个小男孩] 。

Our assumption is that ontological arguments tend to survive literal translations statistically more than non-ontological ones. And Table 4 defines 4 kinds of cross-lingual ontological mapping relations for English and Chinese syntactic arguments. $\boldsymbol{OA_i}$ is defined as a cross-lingual ontological argument here.

Table 4. Cross-lingual Ontological Arguments

	English	Chinese
$\boldsymbol{OA_1}$	NP	NP, BAP, BIP
$\boldsymbol{OA_2}$	AP, AS-AP, DP, PASS-VP, AS-PASS-VP	JP
$\boldsymbol{OA_3}$	VP, TO-VP, WH-TO-VP, VPING, AS-VPING	VP
$\boldsymbol{OA_4}$	SS, AS-IF-SS, WH-SS	SS

Since the English hypothesis generator performs better than its Chinese counterpart, we used the ontological arguments in English hypotheses as heuristic information for Chinese hypothesis generation. We define the cross-lingual ontological argument co-efficiency as follows, where $\boldsymbol{H_E}$ and $\boldsymbol{H_C}$ refer to English and Chinese hypotheses respectively.

$$OAC = \frac{|\{OAs\_in\_H_E\} \cap \{OAs\_in\_H_C\}|}{|\{OAs\_in\_H_E\} \cup \{OAs\_in\_H_C\}|}$$

Now instead of generating only one Chinese hypothesis, we loosen the restrictions on our Chinese hypothesis generator to produce n-best hypotheses, and then choose from the n-best the hypothesis $\boldsymbol{H_C}$ that has the largest cross-lingual ontological co-efficiency with the related English hypothesis, i.e.

$$H_C = Arg\max(OAC_i) =$$

$$Arg\max \frac{|\{OAs\_in\_H_E\} \bigcap \{OAs\_in\_H_{Ci}\}|}{|\{OAs\_in\_H_E\} \bigcup \{OAs\_in\_H_{Ci}\}|}$$

For the Chinese sentence (b) mentioned above, our one-best generator would produce hypothesis NP V VP, while the n-best generator gave three hypotheses, NP V VP, NP V VP NP, and NP V NP. Since the English hypothesis was NP V NP PP, the heuristic method would select NP V NP as Chinese hypothesis at last.

The heuristic method has promoted the token precision of the Chinese hypothesis generator to be 82.7%, and the bilingual hypothesis token precision reached nearly 68%.

4.3 Two-fold MLE Filtering

The two-fold MLE filtering method (Han 2006) is inspired by the theory of diathesis alternations, which are generally regarded as alternative ways that verbs express their arguments. (Han 2006) used the method to filter SCF hypotheses for English verbs.

There are typically two MLE filters employed. For each verb involved, first a common MLE filter is used, but it employs a threshold θ_1 that is much higher than usual, and those SCF hypotheses that satisfy the requirement are accepted. Then, all of the remainder of the hypotheses are checked by another MLE filter seeded with diathesis alternations as heuristic information and equipped with a much lower threshold θ_2. Any hypothesis scf_i left out by the first filter will be accepted if its probability exceeds θ_2 and it is an alternative of an SCF type scf_j that has been accepted by the first filter, which means that $p(scf_i|scf_j, v)>0$ and $scf_j \in \boldsymbol{SCF}_{accepted}$. The filtering process will be performed repeatedly for those unaccepted hypotheses until no more hypotheses can be accepted for the verb.

We modified (Han 2006)'s method a little to adjust it to bilingual SCF acquisition. We used the Chinese SCF diathesis alternations described in (Han 2005) as heuristic information, and the algorithm may be written as follows.

For hypotheses with an English SCF $escf_i$,
1. if $p(escf_i, cscf_i) > \theta_1$,
   accept the hypothesis into set $\boldsymbol{S}$;
2. else
   if $p(escf_i, cscf_i) > \theta_2$,
   and $p(cscf_i|cscf_j) > 0$,
   and $(escf_i, cscf_j) \in \boldsymbol{S}$,
   accept the hypothesis into set $\boldsymbol{S}$;
3. Go to step 1 until $\boldsymbol{S}$ doesn't increase.

Here, $(escf_i, cscf_i)$ is a bilingual SCF hypothesis, and $p(cscf_i|cscf_j) > 0$ means that $cscf_i$ is a diathesis alternative of $cscf_j$.

4.4 Acquisition Results and Evaluation

From the total 650,000 bilingual sentence pairs, we drew out 247,471 sentence pairs with possible parallel predicative verbs, on which the bilingual acquisition experiment was performed.

We manually analysed 50 sentence pairs with typical parallel predicates for each acquired bilingual SCF type that has a probability larger than 0.0005. Against this gold standard, we evaluated the bilingual SCF acquisition results in terms of precision, recall and F-measure of bilingual SCF types. As in SCF acquisition for a single language, precision is the percentage of types that the system proposes correctly, while recall is the percentage of types in the gold standard that the system proposes.

$$\text{Precision} = \frac{|\text{True positives}|}{|\text{True positives}|+|\text{False positives}|}$$

$$\text{Recall} = \frac{|\text{True positives}|}{|\text{True positives}|+|\text{False negatives}|}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here, true positives are correct bilingual SCF types proposed by the system, false positives are incorrect types proposed by the system, and false negatives are correct types that are not proposed by the system.

For the experiment, we empirically set θ_1 to be 0.003, and θ_2 0.0001. Table 5 lists the performances of the two-fold filtering method and a baseline MLE method with a single threshold of 0.0005.

Table 5. Acquisition Performances

Methods	Precision	Recall	F-measure
Baseline	75.38%	65.2%	69.92%
Two-fold	87.6%	81.35%	84.36%

We can see from Table 5 that the two-fold filtering method outperformed the baseline a lot. The precision rate was improved by 12.22%, the recall rate by 16.15%, and F-measure by 14.44%.

5 Analysis on the Acquisition Results

At last, we totally acquired 654 bilingual SCF types for Chinese and English predicative verbs. Against the background corpus, these subcatego-

rization frames are syntactically compatible with probabilities from 0.0001 to 0.0746.

As far as we know, till the present time this is the first study on cross-lingual syntactic subcategorization acquisition for Chinese and English verbs. Therefore, we analyzed further the acquisition results to uncover some linguistic explanation. Our analysis was performed in two aspects, i.e. predicative verbs and bilingual SCF types.

5.1 Predicative Verbs

According to a rough summary of observations on the corpus, the relations between Chinese and English predicative verbs might be classified into three groups.

a. Equivalent predicates: with no syntactic contents gained or lost during the process of translation. Such as, '热爱' vs. 'love', and '购买' vs. 'purchase';

b. Extended predicates: with some syntactic contents gained either on the Chinese or the English side. Such as, '相信' vs. 'believe in', where the English verb 'believe' is extended with a preposition, and '取来' vs. 'fetch', with the Chinese verb '取' complemented by a tendency verb '来';

c. Unparallel predicates: almost unable to be translated into the other language still as predicate. Such as, 'please' vs. '高兴', and 'satisfy' vs. '满意', for which the Chinese verbs are mostly translated into English as 'pleased' and 'satisfied', often annotated as adjectives by Collins' parser, and the English verbs are usually translated into Chinese as '使…高兴' and '令…满意', where the English counterparts are no longer predicates in Chinese.

5.2 Bilingual SCF Types

In facts, the bilingual SCF types are just another kind of representations for relevant syntactic structures of bilingual parallel predicates. Our experiment seems to show that phenomena of cross-lingual subcategorization are quite linguistically comprehensible. The basic bilingual SCF types fall into four classes.

a. Equivalent types: with almost identical syntactic argument structures. Such as,

C_1: NP[我们] V[热爱] NP[祖国] 。
E_1: NP[We] V[love] NP[our motherland] .
C_2: NP[那人] V[走] JP[得极快] 。
E_2: NP[The man] V[went] DP[very fast] .

b. Alternative types: on either the Chinese or the English side, the equivalent SCF being replaced with its diathesis alternation. Such as,

C_1: NP[刘胡兰] BIP[被敌人] V[杀害了] 。
E_1: NP[The enemy] V[murdered] NP[Liu Hulan] .
C_2: NP[老孙] BAP[把墙] V[涂] JP[黑]了。
E_2: NP[Lao Sun] V[painted] NP[the wall] AP[black] .

The Chinese SCF 'NP BIP V' in C_1 is an alternative type of 'NP V NP', while 'NP BAP V JP' in C_2 is an alternative of 'NP V JP NP', and the latter two are equivalent types for their English counterparts.

c. Derivative types: with one or more ontological arguments derived into non-ontological ones. Such as,

C_1: NP[大家] V[要相信] NP[组织] 。
E_1: NP[We] V[should believe] PP[in the organization] .
C_2: NP[这] V[符合] NP[人民的利益] 。
E_2: NP[This] V[complies] PP[with the people's interests] .

The ontological NPs in C_1 and C_2 are derived into non-ontological PPs in E_1 and E_2.

d. Extended types: usually with some more non-ontological arguments realized on either the Chinese or the English side. Such as,

C_1: NP[老爸] V[戒] NP[烟] 了。
E_1: NP[Our dad] V[has given] RP[up] NP[smoking] .
C_2: NP[她] V[取] QP[来了] NP[火石] 。
E_2: NP[She] V[fetched] NP[the firestone] .

The particle argument RP in E_1 and the tendency verbal argument QP in C_2 are used to complement the respective verbs.

And it is obvious that the first two types are closely related to equivalent predicative verbs, while the other two has a lot to do with extended predicative verbs.

6 Conclusion

According to our knowledge, this is the first effort made on cross-lingual subcategorization acquisition for Chinese and English verbs. We, at

first, automatically extracted sentence pairs with possible parallel predicative verbs. Then, by means of heuristic methods, our acquisition experiment established a set of 654 basic bilingual SCF types for the parallel Chinese and English predicates. Further analysis on the experiment results shows that phenomena of syntactic subcategorization are statistically and linguistically compatible.

Proper achievements from this kind of research would surely benefit some natural language processing tasks like machine translation, and cross-lingual information retrieval, etc.

However, there still remains a lot for improvement and adjustment, and approaches that are more complicated still exist theoretically. For instance, English diathesis alternations might as well promote the acquisition performance to a certain degree, bilingual SCF types for individual pairs of parallel verbs are yet to be acquired, and some types unseen by the hypothesis generator might be recalled by integrating semantic verb-classification information into the system.

More essential aspects of our future work will focus on improving the performance of the parallel predicate recognizer and the hypothesis generator, and testing and applying the acquired cross-lingual syntactic subcategorization information in some concrete NLP tasks.

Acknowledgement We are obliged to IRLAB in Harbin Institute of Technology for the dictionary resources they provided us generously. And our great thanks also go to the four reviewers of this paper for their comments, from which we have benefited a lot.

References

Baker, M., 2000. In Other Words: *A Coursebook on Translaton*, Foreign Language Teaching and Research Press, Beijing.

Brent, M., 1993. From Grammar to Lexicon: unsupervised learning of lexical syntax, *Computational Linguistics* 19(3): 243-262.

Cao, Hailong, 2006. *Research on Chinese Syntactic Parsing Based on Lexicalized Statistical Model*, Dissertation for PhD, Harbin Institute of Technology, Harbin.

Chomsky, Noam, 1965. *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chrupala, Grzegorz, 2003. Acquiring Verb Subcategorization from Spanish Corpora, *PhD program "Cognitive Science and Language"*, Universitat de Barcelona.

Collins. M., 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD Dissertation, University of Pennsylvania.

Ellis, R., 2000. *Second language Acquisition*, Shanghai Foreign Language Education Press, Hong Kong.

Gamallo, P., Agustini, A. and Lopes Gabriel P., 2002. Using Co-Composition for Acquiring Syntactic and Semantic Subcategorization, *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, Philadelphia: 34-41.

Han, Xiwu, 2006. *Research on Automatic Acquisition of Chinese Verb Subcategorization*, Dissertation for PhD, Harbin Institute of Technology, Harbin.

Han, Xiwu, Tiejun Zhao and Xingshang Fu, 2006. Improving English Subcategorization Acquisition with Diathesis Alternations as Heuristic Information. *Poster Proceedings of COLING-ACL06*, Sydney:331-336.

Korhonen, Anna, 2001. *Subcategorization Acquisition*, Dissertation for PhD, Trinity Hall University of Cambridge.

Sarkar, A. and Zeman, D., 2000. Automatic Extraction of Subcategorization Frames for Czech, *Proceedings of the 19th International Conference on Computational Linguistics*, Saarbrucken, Germany. Please refer to http://www.sfu.ca/~anoop/papers/pdf/coling0final.pdf

Shulte im Walde, Sabine, 2002. Inducing German Semantic Verb Classes from Purely Syntactic Subcategorization Information, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*: 223-230.

Appendix A Our English Syntactic SCFs

1: IT PASS-V SS //158
2: IT V AP PP TO-VP //??
3: IT V AP TO-VP //??
4: IT V NP NP TO-VP //??
5: IT V NP PP //??
6: IT V NP PP TO-VP //??
7: IT V NP SS //6
8: IT V NP TO-VP //11
9: IT V PP SS //12,108
10: IT V PP TO-VP //10,13
11: IT V PP WH-SS //135
12: IT V RP SS //128
13: IT V SS //107
14: IT V TO-VP //9
15: NP V //22,23
16: NP V AP //1,2
17: NP V AS-IF-SS //159
18: NP V AS-NP //5
19: NP V DP //1,4,160
20: NP V DP PP //119,127

21: NP V IT SS //134
22: NP V NP //24,51,84,123,161
23: NP V NP AP //25,26
24: NP V NP AS-AP //143,147
25: NP V NP AS-NP //29,30
26: NP V NP AS-PASS-VP //162
27: NP V NP AS-VPING //163
28: NP V NP DP //27,28,155
29: NP V NP DP PP //120,152
30: NP V NP NP //37,38,124,144
31: NP V NP NP SS //132
32: NP V NP PASS-VP //58
33: NP V NP PP
//31,39,40,41,42,43,44,45,46,47,48,49,50,56,85,118
34: NP V NP PP PP //122
35: NP V NP PP TO-VP //157
36: NP V NP RP //76
37: NP V NP RP AP //145,146
38: NP V NP RP AS-AP //148
39: NP V NP RP NP //117,125
40: NP V NP RP SS //130
41: NP V NP RP TO-VP //149,150
42: NP V NP SS //52,133
43: NP V NP TO-VP //53,54,55,57
44: NP V NP VP //32,33
45: NP V NP VPING //34,35,36
46: NP V NP WH-SS //59,60,156
47: NP V NP WH-TO-VP //61,62
48: NP V PASS-VP //141
49: NP V PP //63,64,65,69,71,72,73,87,95,96
50: NP V PP PP //91,92,93,94
51: NP V PP PP TO-VP //88
52: NP V PP SS //97,98
53: NP V PP TO-VP //15,66,67,68,99
54: NP V PP VP //153
55: NP V PP WH-SS //89,100,101
56: NP V PP WH-TO-VP //90,102,103
57: NP V RP //74
58: NP V RP AP //137
59: NP V RP DP //126
60: NP V RP NP //76,136
61: NP V RP NP PP //77
62: NP V RP PP //78,140
63: NP V RP PP SS //131
64: NP V RP PP TO-VP //151
65: NP V RP SS //83
66: NP V RP TO-VP //138,139
67: NP V RP VPING //75
68: NP V RP WH-SS //79,80
69: NP V RP WH-TO-VP //81,82
70: NP V SS //104,106,109
71: NP V TO-VP //110,111,112
72: NP V VP //18,142
73: NP V VPING //19,20,21
74: NP V VPING PP //86
75: NP V WH-SS //16,113,114
76: NP V WH-TO-VP //17,115,116
77: SS V //129
78: SS V NP //7
79: SS V PP //14
80: TO-VP V //154
81: TO-VP V NP //8
82: TO-VP V SS //105

Ontologies for a Global Language Infrastructure

Yoshihiko Hayashi
Graduate School of Language and Culture, Osaka University
Toyonaka, 560043 Osaka, Japan
hayashi@lang.osaka-u.ac.jp

Thierry Declerck
DFKI GmbH,
Language Technology Lab
D-66123 Saarbrücken, Germany
declerck@dfki.de

Paul Buitelaar
DFKI GmbH,
Language Technology Lab &
Competence Center Semantic Web
D-66123 Saarbrücken, Germany
paulb@dfki.de

Monica Monachini
Instituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1-56124 Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

Given a situation where human language technologies have been maturing considerably and a rapidly growing range of language data resources being now available, together with natural language processing (NLP) tools/systems, a strong need for a *global language infrastructure* (GLI) is becoming more and more evident, if one wants to ensure re-usability of the resources. A GLI is essentially an open and web-based software platform on which tailored *language services* can be efficiently composed, disseminated and consumed. An infrastructure of this sort is also expected to facilitate further development of language data resources and NLP functionalities. The aims of this paper are twofold: (1) to discuss necessity of ontologies for a GLI, and (2) to draw a high-level configuration of the ontologies, which are integrated into a comprehensive *language service ontology*. To these ends, this paper first explores dimensions of GLI, and then draws a triangular view of a language service, from which necessary ontologies are derived. This paper also examines relevant ongoing international standardization efforts such as LAF, MAF, SynAF, DCR and LMF, and discusses how these frameworks are incorporated into our comprehensive language service ontology. The paper concludes in stressing the need for an international collaboration on the development of a standardized language service ontology.

1 Introduction

With the recent developments of the Semantic Web and progresses of the associated methodologies and standards, demands for an open and distributed infrastructure for sharing language resources and technologies can be addressed now on a new basis (Buitelaar et al., 2003; Calzolari, 2006). In this paper, we call such an infrastructure a *global language infrastructure* (GLI) GLI should accommodate language resources and technologies world-wide. A GLI thus should inherently address multilingual and multicultural issues.

More precisely, a GLI is an open and web-based software platform to which resources can be easily plugged in, and on which tailored *language services* can be efficiently composed, disseminated and consumed. Here a language service simply means a web service whose functionalities are generally related to human language; it can range from simple dictionary access to more complicated linguistic analysis, as well as conversion of linguistic expressions such as translation or paraphrasing.

We can mention the following initiatives/projects as examples of an obvious effort towards such a language infrastructure:

- *CLARIN*[1] is committed to establish an integrated and interoperable research infrastructure of language resources and technology. It aims at addressing the current fragmentation by offering a stable, persistent, accessible and extendable infrastructure that will enable the development of "e-Humanities".
- *Language Grid*[2] provides a language infrastructure on which language services that are useful in intercultural collaboration can be composed, delivered, and utilized. On the Language Grid, existing language data resources, NLP tools/systems and newly created community-based resources can be efficiently and effectively combined (Ishida, 2006). In addition, the Language Grid presents an operation model to address complicated issues associated with intellectual property rights and contracts (Ishida et al., 2008).

These two initiatives share issues of interoperability and reusability of language data resources and NLP tools/systems, even though their primary objectives are totally different. This calls for an opportunity to work out a common strategy for these crucial issues.

With this background, this paper argues that a GLI should be ontology-based, and presents a high-level configuration of the ontologies, which are integrated into a comprehensive *language service ontology*. This paper also examines relevant ongoing international standards, and discusses how these frameworks can be ontologized and incorporated into the comprehensive language service ontology.

2 Dimensions of GLI

2.1 Objectives of GLI

Needs for a language infrastructure have originally emerged from research fields including NLP and a range of e-sciences, which require mining from textual resources. For example, Klein and Potter (2004) presented two use cases; one is a workbench for NLP researchers, and the other is a text-mining tool for e-science researchers who are not necessarily NLP experts.

More recently, CLARIN explicitly targets its users to communities of e-humanities, and tries to offer its services to:

- The different communities of linguists to optimize their models and tools to the benefit of all who are using language material,
- Humanities scholars in the broad sense to facilitate access to language resources and technology, and
- The society as a whole to enable lower thresholds to multicultural and multilingual content.

In contrast, the Language Grid has been launched for providing a language infrastructure for supporting verbal, particularly cross-language, communications that are observed in activities of intercultural collaboration. To achieve this goal, the Language Grid provides an environment in which existing NLP tools/systems and newly created community-based language data resources can be efficiently combined. A number of communication tools are publicized on the project web site.

Here we should remark that (1) the user of a GLI is not necessarily an NLP expert, and (2) not only language data resources but NLP tools/technologies and their useful combinations are involved in a GLI.

2.2 Types of users in GLI

Users, or participants, of a GLI can be classified into the following types:

- A language resource provider who disseminates a language resource or NLP functionality in the form of a language service by wrapping it as a web service,
- A language service composer who composes a composite web service by combining atomic language services, and
- A language service end user who simply consumes a language service.

From a language infrastructure perspective, it is of crucial importance to provide useful support for a language resource provider in creating the wrap-

[1] http://www.clarin.eu/
[2] http://langrid.nict.go.jp/

pers, and for language service composers in authoring composite language services. To these ends, a standardized framework for describing language data resources and NLP tools/systems is strongly required (Hayashi, 2007).

2.3 Technical ingredients of GLI

As implied from the discussions so far, technical ingredients of a GLI are: (1) NLP tools/systems ranging from dictionary access systems and linguistic analyzers to machine translation systems, and (2) language data resources, such as lexicons or corpora. In addition to these, a GLI has to be aware of abstract linguistic objects such as linguistic expression, linguistic annotation or even linguistic meaning, because these types of abstract objects comprise the data to/from NLP tools/systems, as well as content of language data resources.

3 Ontologies for a GLI

3.1 Necessity of a comprehensive ontology

In principle, most of the existing language data resources and NLP tools/systems have been created independently, resulting in a situation where data format, annotation scheme, access method and other features are all idiosyncratic. This obviously will be a burden for establishing a GLI which ensures interoperability and reusability of language data resources and NLP tools/systems. To address this issue, standardization is inevitable: standardized APIs are necessary for NLP tools/systems; standardized data semantics as well as data format are required for language data resources. In addition and importantly, these standards should be designed based on a comprehensive shared ontology which covers all possible elements of a GLI.

3.2 Triangular view of a language service

In order to facilitate the development of a comprehensive ontology, it should be divided into appropriate sub-ontologies, each covering a grouped set of elements. Figure 1 shows a triangular view of a language service. Note that a language service is provided by a *language process*, not solely by language data or linguistic objects. Therefore language processes should be placed at the vertex of the triangle. A language process, in general, processes a linguistic expression which may or may not be linguistically annotated. We denote abstract objects such as linguistic expression or linguistic annotation as *linguistic object*. Linguistic objects may comprise a *language data* resource such as a corpus or lexicon; hence it would be utilized by a language process. This triangular view of a language service gives us a foundation on which necessary sub-ontologies are developed.

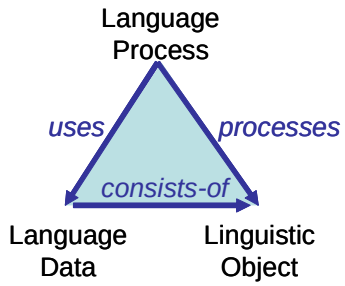


Figure 1: Triangular view of a language service.

3.3 Top-level of the language service ontology

Figure 2[3] illustrates the top-level of the language service ontology that is configured according to the language service triangle depicted in Fig 1. Each box in the figure denotes a top-level class in the ontology, which is defined in further detail by a sub-ontology. Among these top concepts, **`LanguageService`** is the top-most concept. As discussed with the language service triangle, a language service is provided by **`LanguageProcessingResource`** which takes **`LinguisticExpression`** as input/output and uses **`LanguageDataResource`**. Note that a language data resource does not provide a language service by itself; it is always used through an access mechanism which is an instance of some sub-class of the processing resource class.

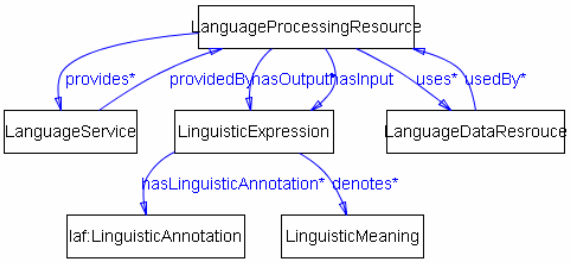


Figure 2: Top-level of the language service ontology.

[3] All the figures (except Fig.1) were produced with the OntoViz plugin of the Protégé ontology editor.

In further detailing the sub-ontologies, we believe it to be important to incorporate related international standards.In this sense, we have been looking at frameworks for linguistic annotation and lexicon modeling that have been discussed in international standardization bodies. The frameworks for linguistic annotation are incorporated into our ontology not only for specifying the input/output data of NLP tools, but also for defining the content of corpora. On the other hand, the framework for lexicon modeling is introduced to have a formal foundation for developing a taxonomy of lexicon classes, which are obviously subclasses of the language data resource class.

4 Ontology for Linguistic Annotations

Figure 2 also depicts an ontological configuration for abstract linguistic objects such as linguistic expression, linguistic meaning and linguistic annotation. It says: (1) a linguistic expression (**`LinguisticExpression`**) in a language denotes some meaning (**`LinguisticMeaning`**), even if it is not explicitly represented, (2) a linguistic expression should be the input or the output of a language process, and (3) a linguistic expression can be multiply annotated (**`LinguisticAnnotation`**). The last point is of crucial importance, because any framework for linguistic annotation has to be able to accommodate multiply layered annotations, given the possibility that the target linguistic expression would be annotated by more than one analyzer, each of which possibly doing its job on a different linguistic level. Among the linguistic objects, ontological configuration for the linguistic annotation should be most carefully designed with respect to the interoperability and reusability of language data resources and NLP tools, because the data to/from a linguistic analyzer, as well as the content of a language data resource should be represented as linguistic annotation.

Frameworks that are necessary for standardized linguistic annotations have been actively developed and disseminated by the ISO TC37/SC4 [4] committee; these include LAF (Linguistic Annotation Framework) (Ide and Romary, 2006), MAF for morphosyntactic annotation (Clément and de la Clergerie, 2005), SynAF for syntactic annotation (Declerck, 2006), and others. Among these, the LAF is the most general *umbrella* framework, and the other frameworks inherit the basic properties of LAF. As these frameworks have not been defined in the form of an ontology, we decided to *ontologize* these frameworks and incorporate them into the language service ontology. Here to ontologize simply means to give OWL (Web Ontology Language) (McGuinness and Harmelen, 2004) specifications to relevant parts of the framework.

Figure 3 illustrates a high-level configuration of the sub-ontology for linguistic annotations. This configuration corresponds to the LAF framework. As shown in the figure, a linguistic annotation has a start position and an end position for designating the span of annotation in the target linguistic expression[5]. This allows us to implement so-called stand-off annotation, and hence enables multiple annotations on the same data set. It also accommodates a feature structure for representing the annotation content.

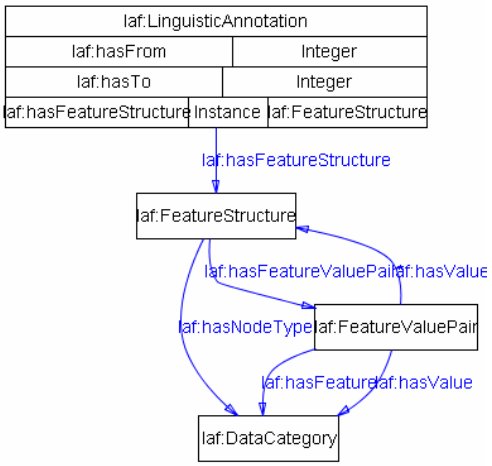


Figure 3: Configuration for LAF.

As noted in (Declerck et al., 2007), the LAF does not provide specifications for content categories; instead it includes a DCR (Data Category Registry) (Wright, 2004) that contains pre-defined data elements and schemas that may be used in annotations. Current configuration of the data categories does not induce taxonomical structure. Nevertheless the linguistic annotation class should be further organized into sub-classes based on which data categories should be included.

Figure 4 summarizes the ontological configuration for MAF and SynAF, introducing classes for

[4] http://www.tc37sc4.org/

[5] In LAF, this is called *primary data*.

segment (**SegmentAnnotation**), syntactic constituent (**SyntacticAnnotation**), and dependency relation (**DependencyRelation**). Note that these classes have been explicitly introduced, although these, in principle, should be represented with the feature structures. Although it is not depicted in the figure, the feature structure for representing morpho-syntactic annotation attached to a segment should be restricted to only include MAF conformant data categories. A similar story should apply to SynAF. As proposed in (Declerck, 2006), SynAF is designed to be able to represent two syntactic properties of a human language: *constituency* and *dependency*. Therefore the syntactic annotation class should be defined to have a specialized feature structure whose node type is restricted to the categories defined in the data category sub-profiles for constituency relation or dependency relation.

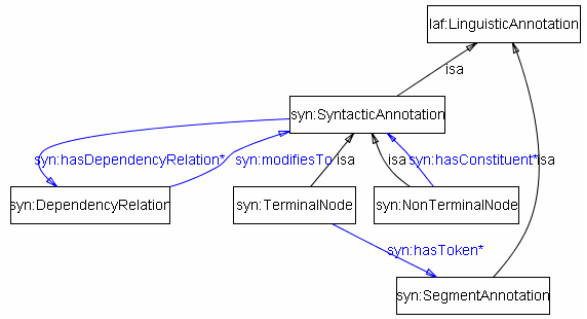


Figure 4: Configuration for MAF and SynAF.

With the ontology described so far, any linguistic expression in the proposing language service ontology can be typed according to the type of linguistic annotation it has. This type information can be effectively utilized in dynamic composition of composite services, in which checking of the input/output constraints given in the meta-description of a processing resource is necessary.

5 Ontology for Lexicons

The class for language data resource (**LanguageDataResource**) is currently organized by subclasses for corpus (**Corpus**) and lexicon (**Lexicon**). The corpus class can be further organized into subclasses according to the type of content, where type can be defined by the type of annotation of the content. Thus we can have an interrelation between the corpus ontology and the linguistic annotation ontology.

Similarly but not identically, the lexicon class should be organized into subclasses by the type of lexical content, and the type should be defined based on the features of a lexical entry in the target lexicon. Here we have an opportunity to incorporate ongoing standardization work in lexicon modeling into our language service ontology. To do this, we have first ontologized parts of the LMF (Lexical Markup Framework) (Francopoulo et al., 2006) which is also in the process of standardization by ISO TC37/SC4, and then connected these with the lexicon class taxonomy.

The ultimate goal of LMF, as stated in (ISO DIS 24613:2007), is to create a modular structure that will facilitate true content interoperability across all aspects of electronic lexical resources. The modular structure of LMF consists of a core package and a number of extensions for modeling a range of machine readable dictionaries (MRDs), and NLP lexicons. These LMF extensions are expressed by extending the LMF core package, encouraging us to ontologize them by organizing the classes defined in the core package as subclasses of the top LMF class.

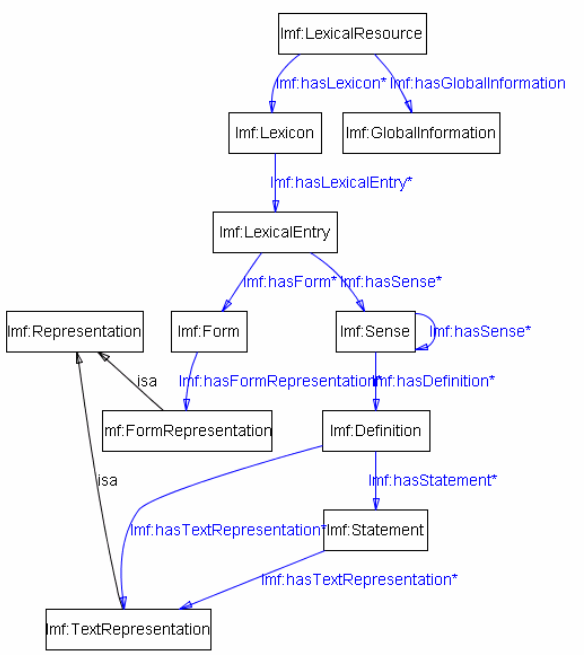


Figure 5: Configuration for LMF core model.

Figure 5 illustrates the ontological configuration for the LMF core model, while Figure 6 shows a part of the LMF NLP Semantics extension, which is associated in particular with the lexical semantic

notions of the extension. As seen in these examples, ontologization of LMF is rather straightforward.

Then the questions should be:

- How can we define the taxonomy of the lexicon while referring to the ontologized LMF?

Figure 7 shows a taxonomy of the lexicon class, stating that each of the lexicon subclasses is defined in terms of the type of the lexical entries defined in the ontologized LMF. For example, **`BilingualDicctionary`**, a sub-class of **`MRD`**, is defined by **`hasLexicalEntry`** property whose

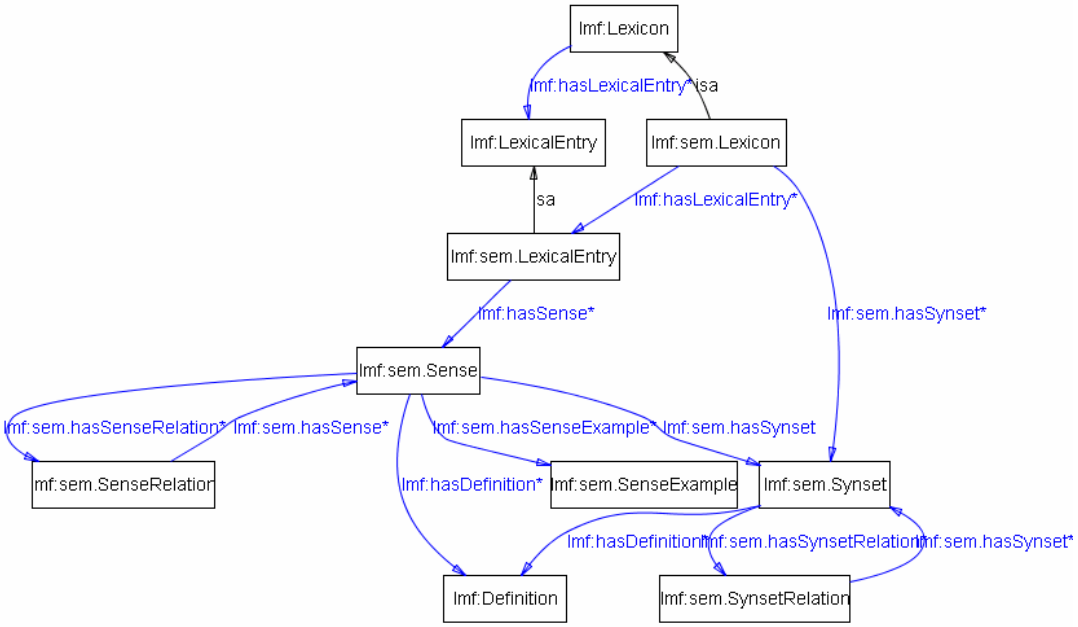


Figure 6: Configuration for LMF NLP Semantics.

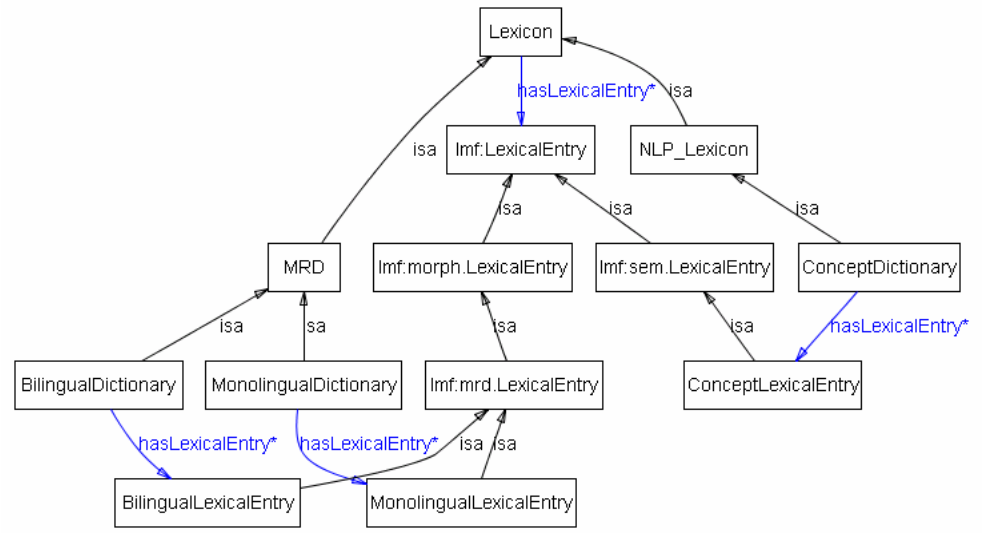


Figure 7: Taxonomy of lexicon class.

- How can we define a class for lexicon access, which is a sub-class of the processing resource (**`LexiconAccessor`**), while referring to the ontologized LMF?

range is restricted to **`BilingualLexicalEntry`**, which, in turn, is a descendant class of **`LexicalEntry`**. In order to incorporate some new type of lexicon, we have to first introduce a new sub-class, then appropriately place it some-

where in the lexical entry class in the ontologized LMF and finally relate it to the lexicon taxonomy.

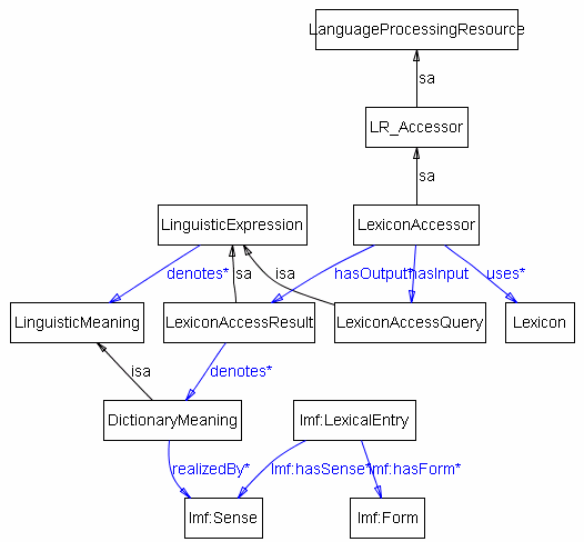


Figure 8: Configuration of lexicon accessor class.

Figure 8 summarizes the ontological definition for the lexicon accessor class; its input is restricted to a sub-class of the linguistic expression class (**`LexiconAccessQuery`**), whereas the output is restricted to **`LexiconAccessResult`** which is also a sub-class of the linguistic expression class. The former is defined to have properties for query conditions, while the latter is restricted to denote an instance of **`DictionaryMeaning`**, which is a sub-class of **`Meaning`**. Note that the dictionary meaning would be **`realizedBy`** an instance of the **`Sense`** class in the ontologized LMF. Here we have an explicit interrelation between the part of language service ontology with the LMF ontology. Note also that the **`Sense`** instance is associated with an instance of **`LexicalEntry`** class, and the associated **`Form`** instance should match with the linguistic expression given in the input query to the lexicon accessor. A *deep* constraint like this, however, is unfortunately beyond the representivity of the OWL formalism, hence not explicitly encoded. To encode such a deep constraint, the notion of *process* have to be introduced with a framework (e.g. SWRL) (Horrocks, et al., 2003) for expressing complicated logical relationships.

6 Related Work

Klein and Potter (2004) sketch an ontology for NLP services with OWL-S specifications. Their proposal unfortunately did not include ontologies for abstract linguistic objects such as linguistic annotations. Hayashi (2007) proposed a linguistic service ontology in the context of the Language Grid. Although it discussed a taxonomy for NLP tools, it did not present any details on the linguistic annotation and lexicon modeling.

LT World (Jörg and Uszkoreit, 2005) is a comprehensive knowledge portal for language technologies. One of the unique features of LT World is that it is based on a multi-dimensional ontology. For example, it classifies language technologies into such dimensions as: application, linguality, languages, technologies, linguistic area, and linguistic approach. This part of the ontology could be incorporated into our ontology especially for specifying the language processing resources.

Several relevant frameworks around language data resources have been actively developed by ISO TC37/SC4. As noted in this paper, we will carefully observe the activities, and incorporate the results as much as possible into our language service ontology. Among these, future development of the DCR will be of importance. That is, by developing an ontology for linguistic categories on top of the basic DCR data categories, we will have an opportunity to explicitly define relations among the data categories in our language service ontology. In this regard, our approach to the ontology for linguistic categories is in some degree different from the one taken by GOLD (Farrar and Langendoen, 2003), where not only linguistic categories but complex relations among them are fundamentally defined within the central ontology.

7 Concluding Remarks

A *global language infrastructure* (GLI) an open and web-based software platform on which tailored language services can be efficiently composed, disseminated and consumed. Given the increasingly realistic scenario in which language data resources and NLP tools/systems will be ubiquitous on the web, a comprehensive ontology (*language service ontology*) for describing these elements will be vital in addressing such issues in interoperability and reusability.

In this paper, we have examined a triangular view of a language service, which consists of language processing, language data, and linguistic objects. Based on this definition, we have pre-

sented a top-level ontology configuration along with an essential set of sub-ontologies; these include ontologies for processing resources, language data resources, linguistic annotations, and lexicons. Among these, the ontologies for linguistic annotations and lexicons have been substantially detailed while referring to the ISO frameworks LAF, MAF, SynAF, DCR, and LMF. In doing so, we ontologized an essential part of these frameworks, and incorporated them into our comprehensive language service ontology.

We strongly believe that although the results presented in this paper are still preliminary, the resulting language service ontology will be essential in defining an ontology-based GLI. Obviously, we still have to provide further detail for the presented sub-ontologies by looking at concrete language data resources and NLP tools/systems for a range of human languages. In parallel, we will need to develop an approach for handling any differences in desired expressiveness inherent to the objective of a GLI; e.g., a language research infrastructure may require precise linguistic descriptions, while an infrastructure for NLP applications might demand more coarse-grained linguistic descriptions, while focusing rather on detailed communicative aspects.

To conclude, in reaching an ontology-based GLI, we will need to establish a community of experts from a range of relevant research areas and human languages. We sincerely hope that this paper will contribute to the initiate such an initiative.

Acknowledgment

This research was in part supported by "R&D promotion scheme funding international joint research" promoted by NICT, Japan.

References

Paul Buitelaar, Thierry Declerck, Nicoletta Calzolari, and Alessandro Lenci. 2003. Language Resources and the Semantic Web. In: *Proc. of ELS-NET/ENABLER workshop*.

Nicoletta Calzolari. 2006. Community Culture in Language Resources - An International Perspective. In : *Proc. of LREC2006 Workshop Towards a Research Infrastructure for Language Resources.*

Lionel Clément, and Éric Villemonte de la Clergerie. 2005. MAF: a morphosyntactic annotation framework. In: *Proc. of LTC2005*.

Thierry Declerck. 2006. SynAF: Towards a Standard for Syntactic Annotation. In: *Proc. of LREC2006*, pp.229-233

Thierry Declerck, Nancy Ide, and Thorsten Trippel. 2008. Interoperable Language Resources. In: *Sprache und Datenverarbeitung (International Journal for Language Data Processing)*, to appear.

Scott Farrar, and Terry Langendoen. 2003. A Linguistic Ontology for the Semantic Web. *Glot International*, Vol.7, pp.97-100.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In: *Proc. of LREC2006*, pp.233-236.

Yoshihiko Hayashi. 2007. Conceptual Framework of an Upper Ontology for Describing Linguistic Services. In: Toru Ishida, Susan R. Fussell, Piek T. J. M. Vossen (Eds.): *Intercultural Collaboration*, LNCS 4568, Springer-Verlag, pp.31-45.

Ian Horrocks, et al. 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. http://www.w3.org/Submission/SWRL/

Nancy Ide, and Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. In: *Proc. of LREC2006*, pp.225-228.

Toru Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. In: *Proc. of SAINT2006*, pp.96-100.

Toru Ishida, et al. 2008. A Non-Profit Operation Model for the Language Grid. In: *Proc. of ICGL2008*, to appear.

ISO DIS 24613:2007. 2007. Language resource management -Lexical markup framework (LMF), Rev.14.

Brigitte Jörg, and Hans Uszkoreit. 2005. The Ontology-based Architecture of LT World, a Comprehensive Web Information System for a Science and Technology Discipline. In: *Leitbild Informationskompetenz: Positionen - Praxis - Perspektiven im europäischen Wissensmarkt*.

Evan Klein, and Stephen Potter. 2004. An Ontology for NLP Services. In: *Proc. of LREC Workshop on a Registry of Linguistic Data Categories within an Integrated Language Resource Repository Area*.

Deborah L. McGuinness, and Frank van Harmelen. 2004. OWL Web Ontology Language Overview. http://www.w3.org/TR/owl-features/

Sue Ellen Wright. 2004. A Global Data Category Registry for Interoperable Language Resources. In: P*roc. of LREC2004*, pp.123-126.

Global Interoperability: How Can We Get There?

Nancy Ide
Department of Computer Science
Vassar College
USA
ide@cs.vassar.edu

Over the past two decades, the notion of "interoperability" has meant different things for the natural language processing (NLP) community. Twenty years ago, NLP researchers had many of the same desiderata as most computer users: software that could run on any platform, and data in formats that could be immediately input to software for which it was not originally designed without substantial programming effort to transduce it. Platform independence has been largely achieved at this time, and, due to continued efforts since the late 1980's, we are beginning to see some convergence in the use of common, reusable data formats for language resources. In recent years, efforts toward interoperability have focused on harmonizing linguistic categories—including everything from part of speech categories to semantic and discourse-level information. The universal use of common annotation categories has obvious advantages, especially for linguistic information that requires a significant amount of manual intervention to produce: it would enable the distribution of annotation effort, and, perhaps more crucially, it will enable studying interactions across linguistic levels, which is undoubtedly the next major step to improve NLP. The development of common linguistic descriptors poses the major challenge to the NLP community today, and which is the focus of many of the papers at this conference. But as we continue to work toward this interim goal, it is useful to step back and consider the longer-term goals we are aiming to reach, and the shape that a truly global language processing infrastructure may eventually take.

Our ultimate vision of language understanding by machine is, at this point, the stuff of science fiction, where computers are indistinguishable from humans (except, perhaps, in their super-human ability to understand or generate even newly encountered languages). We can imagine that this would require a processing system with a configuration similar to that of the human brain, consisting of billions of neurons knitted together in a complex network of connections. Although this kind of language network is likely to remain far beyond our capabilities for some time to come, we can envision an interim solution that takes a step in that direction: a fully interlinked, globally distributed network of multilingual language resources and language processing tools that are accessible for dynamic NLP, such that as soon as language data is accessible, either in spoken or written form, it is immediately rendered into some representation that enables retrieval of the information it contains. Very faint glimmers of this kind of network are in existence today, in the Global WordNet and projects such as Kyoto University's Language Grid. But it is humbling to consider how far we are even from this interim vision. What will it take for us to get from here to there?

My presentation will outline some possible scenarios for global interoperability in the future, and consider the steps we are currently taking as well as those we may take to create a global infrastructure for language processing. It will, hopefully, provide a context for discussion both among the members of the NLP community attending this conference concerning the directions and goals of NLP research for the foreseeable future.

A Non-Profit Operation Model for the Language Grid

Toru Ishida[1,2], Akiyo Nadamoto[2], Yohei Murakami[2], Rieko Inaba[2], Tomohiro Shigenobu[2],
Shigeo Matsubara[1,2], Hiromitsu Hattori[1], Yoko Kubota[1],
Takao Nakaguchi[3], and Eri Tsunokawa[3]

[1] Department of Informatics, Kyoto University, Japan
{ishida, matsubara, hatto, yoko}@i.kyoto-u.ac.jp
[2] Language Grid Project, National Institute of Information and Communications Technology, Japan
{nadamoto, yohei, rieko.inaba, shigenobu}@nict.go.jp
[3] NTT Advanced Technology Corporation, Japan
{takao.nakaguchi, eri.tsunokawa}@ntt-at.co.jp

Abstract

The Language Grid is an initiative to build an infrastructure that allows end users to create new language services for their intercultural / multilingual activities. To this end, language resources (including data and programs) are wrapped to form Web services that users can combine easily to realize workflows that suit their own activities. There are four types of stakeholders in the Language Grid: *Language Resource Provider*, *Computation Resource Provider*, *Language Service User*, and *Language Grid Operator* (who coordinates the other stakeholders). Though there can be various operation models for the Language Grid, we propose a non-profit operation model in this paper. This model limits the usage of language services solely to non-profit operations, tries to match the incentives of stakeholders, and manages various problems associated with intellectual property rights, user privacy, and operation costs.

1 Introduction

In 2002, we conducted a six month experiment called the *Intercultural Collaboration Experiment (ICE2002)* to develop open source software with Chinese, Korean and Malaysian colleagues (Nomura *et al.*, 2003). We thought that machine translation would be useful in facilitating intercultural activities. We gathered machine translators to cover five languages: Chinese, Japanese, Korean, Malay and English. More than forty students and faculty members joined this experiment during which they discovered the accessibility and usability of machine translators.

Even though there are many language resources (both data and programs) on the Internet (Choukri, 2004), many ongoing intercultural collaboration activities are still lacking multilingual support. We have been working with nonprofit organizations to investigate the requirements placed on the Language Grid. One NPO[1] have been creating a "universal playground" for kids around the world. To plan activities among Japan, Korea, Kenya and Austria, volunteer facilitators had to use their own dictionaries for performing translations. Another NPO[2] assisting foreign patients must use multilingual parallel texts for medical support. However, most end users have no way of employing the extant language resources, because of complex intellectual property rights, non-standardized technical interfaces, and so on. If technologies were available that could provide software to coordinate stakeholders, to combine language resources, and to create language services for end users, it is likely that people will start to use language resources in daily life (Ishida *et al.*, 2007).

The Language Grid project (Ishida, 2003) aims at wrapping language resources as Web services so as to make it easy to manage intellectual property. *Language Grid Users* can take three different roles: *Language Resource Provider*, *Computation Resource Provider*, and *Language Service User*. To conclude agreements between Language Grid

[1] NPO Pangaea. http://www.pangaean.org/.
[2] Kyoto Center for Multicultural Society http://www.tabunka.jp/kyoto/.

Operator and Language Grid User, we need to discuss operation models of the Language Grid[3].

Though various operation models are possible for the Language Grid, this paper proposes a *non-profit operation model*, which limits the usage of language services to non-profit activities. The model is designed to fulfill the requirements of all stakeholders. Because the model is intended for non-profit use, universities or research institutes should be able to operate the Language Grid, and thus the cost of operation should not be too large. The rest of paper describes the service layer, stakeholders, requirements, and proposed operation model of the Language Grid.

2 Service Layer

2.1 Service Layer of the Language Grid

As shown in Figure 1, the Language Grid consists of four service layers.

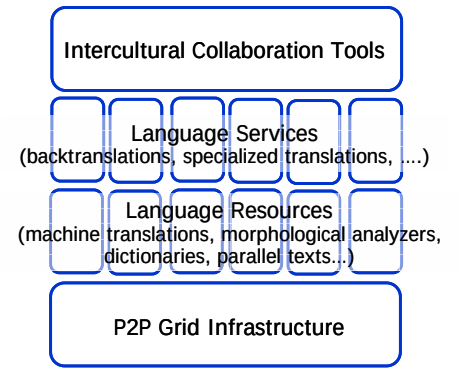


Figure 1. Service Layer

a) Intercultural Collaboration Tools
Collaboration tools are developed using language services explained below. Though the Language Grid provides several tools, new tools can be developed by Language Service Users, and also existing tools can be multilingualized.

b) Language Services
Various language services will be available by combining existing language resources. We already implement Web service workflows including *back translations*[4] and *domain specific translations*. Language Service Users can easily add new language services to the Language Grid by themselves.

c) Language Resources
Various language resources will be provided as atomic Web services on a standardized interface. Language Resource Providers can easily add new language resources to the Language Grid, and access the usage statistics of their resources.

d) P2P Grid Infrastructure
This infrastructure organizes multiple servers on the Internet to fulfill end users' requests. Computation Resource Providers can add their servers to the P2P Grid, and access the usage statistics of their resources.

2.2 Intercultural Collaboration Tools

The Language Grid provides several multilingual communication tools that combine community dictionaries and machine translators.

a) Langrid Input
A multilingual input interface for existing collaboration tools. As shown in Figure 2, input texts are translated, in real time, into various languages and sent to collaboration tools. Language Service Users can multilingualize existing tools by attaching Langrid Input interfaces to them. Users can also edit dictionaries of community vocabularies[5].

b) Langrid Chat
A chat tool with multilingual translations. Users can read and write messages in their first language. Langrid Input is used as a text input interface.

c) Langrid Blackboard
A tool for summarizing and sharing multilingual information. Users can create cards in their first languages and post them in a shared workspace. The texts on the cards are immedi-

[3] Besides language and computation resource providers, we need to consider providers of Web service wrappers, Web service workflows, and various multilingual collaboration tools that use the Language Grid. To simplify the discussions, we do not explicitly mention those stakeholders in this paper. We assume that the above software, as well as Language Grid software that run on Language Grid servers, are already licensed to the Language Grid Users.

[4] To allow users to check the accuracy of translation, collaboration tools often provide *back translation*, which translates the translated sentences into the original input language.

[5] Natural Disaster Youth Summit Committee has been using NGO iEARN's BBS with the Langrid Input. http://ndys.jearn.jp/

ately translated into different languages. This tool is useful for international meetings. Langrid Input is used as a text input interface[6].

Figure 2. Langrid Input

2.3 Language Services

Among existing research studies, EuroWordNet (Vossen, 1998) and Global WordNet Grid (Fellbaum and Vossen, 2007) are pioneer works on connecting dictionaries in different languages based on word semantics. The Language Grid, however, is a trial to build an infrastructure that can combine language services by combining the incentives of stakeholders.

The Language Grid uses WS-BPEL to describe workflows, and uses the WS-BPEL engine to execute them; the engine sequentially invokes Web services as specified in the workflow. The following language services are available in the Language Grid:

a) Atomic service: A Web service that corresponds to individual language resource. For example, bilingual dictionaries, parallel texts, morphological analyzers, and machine translators.

b) Composite service: An advanced service described by a workflow that combines several atomic services. For example, multiple dictionary search, domain specific translations, and back translations.

Figure 3 shows a typical composite service: several atomic services are combined to create domain specific translation.

[6] Tools are available from http://langrid.nict.go.jp/.

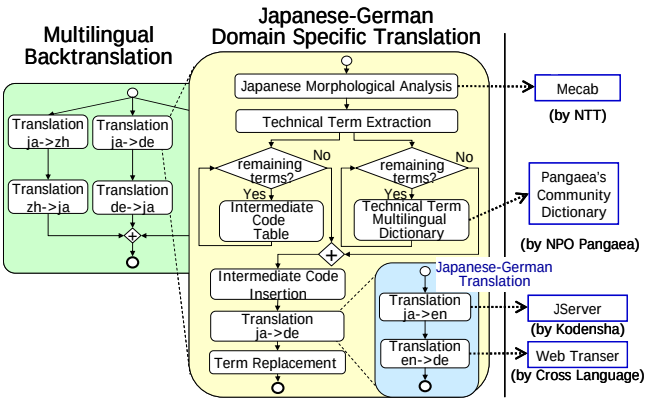


Figure 3. Example of Composite Service

2.4 Language Resources

The Language Grid can grow through the voluntary efforts of Language Grid Users. The more users provide resources, the more they appreciate the benefits of the resources. However, to create the initial seed, we need machine translators and morphological analyzers for both European and Asian languages.

To use language resources, we need to wrap them as Web services. Language Service Users can register Web services and share them via the Language Grid. For this purpose, standardization of access entry is quite important (Calzolari 2002). We started working on Language Service Ontology, which standardizes the interfaces for wrapping language resources (Hayashi and Ishida, 2006) (Hayashi *et al.*, 2008).

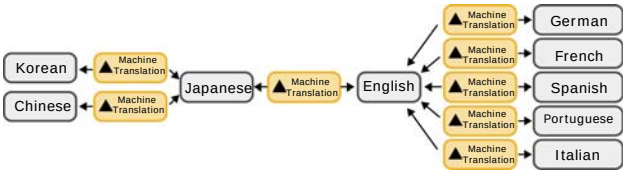


Figure 4. Available Machine Translators

To increase the flexibility of composing language services, a wide variety of language resources must be registered. However, available language resources are often limited. Figure 4 shows the machine translators available in the current Language Grid. English is often the hub of translation, and that makes back translation complex. If we use Japanese-German back translation, we have to combine four translators provided by

different organizations[7]. We have observed how the quality of translation effects communication (Yamashita and Ishida, 2006). The findings have contributed to the development of new technologies to coordinate multiple machine translators.

2.5 P2P Grid Infrastructure

The P2P Grid Infrastructure is aimed at connecting servers around the world. As shown in Figure 5, the P2P Grid consists of two kinds of servers: *core nodes* and *service nodes*.

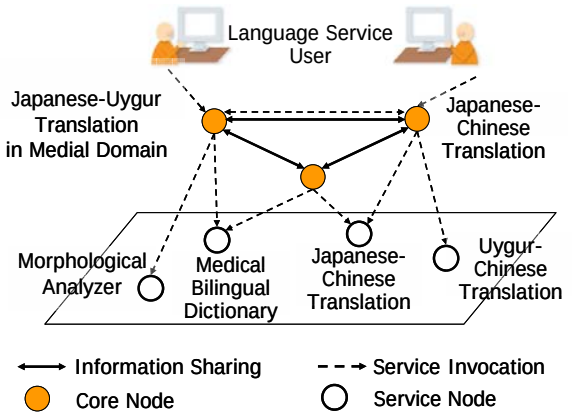


Figure 5. P2P Grid Infrastructure

Core nodes manage registered language services. They provide the search functions used by language services. Based on users' requests, core nodes invoke composite Web services by executing workflows. Information of language resources is shared among all core nodes. As a result, same services are provided, regardless of which core node is accessed by the user. The core nodes also control access to service nodes.

The service nodes deploy various language resources as Web services. The Language Grid sets basic authentication functions on the service nodes. Therefore, access from nodes other than core nodes is not accepted by service nodes.

There already exist several efforts to combine language processing programs: Heart of Gold (Callmeier et al., 2004), and UIMA (Ferrucci and Lally, 2004). Though there are similarities between Heart of Gold, UIMA and the Language Grid, their focuses are orthogonal. Heart of Gold and UIMA aim at allowing language processing programs with variable interfaces to share data, while the Language Grid focuses on managing the intellectual property associated with language resources (both data and programs) via the Web service architecture. We started bridging Heart of Gold and the Language Grid and will apply the results to combine UIMA and the Language Grid.

[7] NPO Pangaea has been used Japanese-Korean and English-German translations with their own dictionaries, but does not use Japanese-German translation because of quality of back translation.

3 Stakeholders and Requirements

3.1 Stakeholders

As shown in Figure 6, the term "Language Grid User" means the three types of stakeholders.

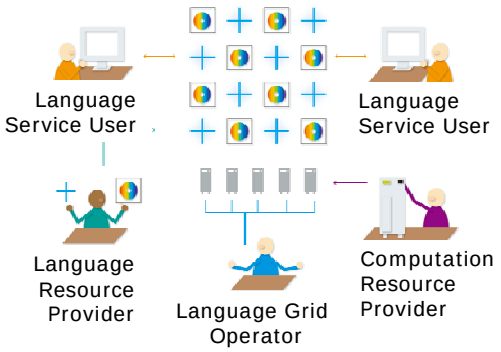


Figure 6. Stakeholders

Language Resource Providers register language resources, Computation Resource Providers register servers with the Language Grid, and Language Service Users use registered language and computation resources. Language Grid Operator concludes agreements with the Language Grid Users and manages language and computing resources.

3.2 Requirements

To design a non-profit operation model, we should first collect the requirements of the stakeholders. University laboratories and research institutes are expected to join the Language Grid as Language and Computation Resource Providers, and various NPOs, NGOs and public sectors are expected to join as Language Service Users. To create a non-profit operation model, we need to understand the following situations and requirements from stakeholders.

- Machine translators are often developed and operated by for-profit companies, and are provided for profit. However, if the application area of the Language Grid does not conflict with an already

existing business market, we can collaborate with those companies and receive a substantial discount on prices. One solution is that universities, research institutes or large NGOs voluntarily buy translation services and provide them to the Language Grid without any charge[8].

- Morphological analyzers and other language processing programs are often developed by research institutes or universities. In many cases, researchers can provide their resources without any charge for research purposes. Even if the goal is not for research, if their use can be restricted to non-profit, researchers often agree to provide their resources. For profit tools, however, are seldom free and contracts cannot be concluded uniformly.
- Multilingual dictionaries and multilingual parallel texts may or may not be free. Even for non-profit use, if the resources are already being sold, difficult problems exist with regard to the distribution of those resources without charge. Since the Language Grid is based on Web services, however, there is a chance of making those resources freely available by setting an upper limit to daily access number.
- Most language resource providers are willing to disseminate the fruit of their research and development, expecting their resources to be widely used and to contribute to intercultural activities. However, they require that the provided resources are used properly following the agreement they signed.

The Language Grid is based on Web service technologies to combine software on the Internet. What Language Grid offers is, however, not only composite Web services but also an infrastructure wherein stakeholders can provide and/or use language resources by mutual consent, understanding and solving the intellectual property issue in each case. To meet this goal, as described in detail in Section 4, we need a *Language Grid Service Manager* that allows stakeholders to monitor the grid and to set and confirm information necessary for their participation.

We should also consider that Language Grid Operators of non-profit models are often laboratories of university or research institutes. We should make the costs of operation as small as possible. Those Operators may not be able to handle personal information, nor invoices or transactions for the fee-based usage of language resources.

4 Non-Profit Operation Model

4.1 Language Grid Users

A Language Grid User could be more than one type of stakeholder at the same time. For example, a university that is providing a morphological analyzer might use a machine translation engine provided by another stakeholder. In this case, the university is a Language Service User as well as a Language Resource Provider. Therefore, we call all three types of stakeholders Language Grid Users. Language Grid Users are required to conclude an agreement with the Language Grid Operator.

The term "non-profit use" in this paper means the use by individuals, the use by public agencies or nonprofit organizations for their main businesses or for research, and the use by for-profit organizations for social contribution. Note that we do not exclude for-profit contracts concluded outside of the Language Grid to provide language resources with charge. What we strictly exclude is to use language resources provided via the Language Grid for profit purposes.

Language Grid Users are, in most cases, expected to be public agencies or non-profit organizations. However, private enterprises are eligible to be Language Grid Users if they use the Language Grid for voluntary social action programs, or they behave just as Language Resource Providers[9]. Individuals can also be Language Grid Users. For example, when a researcher working for a private enterprise provides language resources for which he/she holds the copyright, the person is eligible to become a Language Grid User.

4.2 Language Resource Providers

The Language Grid Operator makes language resources accessible via the Language Grid for non-profit use. The Operator can, with Providers' consent, select or change the computation resources in which Providers' language resources are deployed.

Language Resource Providers can set a copyright notice and/or any licensing policy informa-

[8] At the starting point of Language Grid operation, we negotiated with commercial companies to provide the machine translators illustrated in Figure 4.

[9] NTT Communication Science Laboratory provides Mecab, a Japanese morphological analyzer.

tion in the *Profile* of their language resources using the Language Grid Service Manager. Providers can also enter a URL to a Web site that describes the license.

The Language Service Users can view the profile. Furthermore, when their language service is used, the Language Grid sends copyright and license information to the Users, so that they can display the copyright and license information on their collaboration tools.

Language Resource Providers can monitor statistics on how their resources are being used. Providers can monitor the following information.

- Total number of accesses of each user to provided language services per year, month, and day.
- The registration information (name of organization, responsible person, email address, URL, etc.) of users of provided language services

Language Resource Providers can control access to the language resources as follows.

- Language Resource Providers can permit or deny access by certain users by setting access rights. When a Provider registers a language resource, its default setting is that no Language Service User is allowed to access it. The Provider then configures the settings to give permission to users. Language Resource Providers can also set the effective period of the access right.
- Language Resource Providers can restrict the total number of accesses per year, month, and day to their resource to prevent inappropriate usage. They can also set the effective period for each access frequency restriction.
- Language Resource Providers can restrict the amount of data transferred per year, month, and day to prevent the download of all data in their resources. Furthermore, they can set the effective period for data transfer. The maximum volume of data transfer per access can be set as well.

If, for some reason, Language Resource Providers want to provide their resources as paid services, they should negotiate directly with Language Service Users[10]. The Operator will not be involved in any contracts wherein language resources are provided with charge.

If Language Resource Providers need any detailed log information, they should negotiate directly with Language Service Users for obtaining such information. To protect the privacy of Language Service Users, Language Grid Operator will not store or collect any log information other than statistics.

If necessary, Language Resource Providers can suspend the use of their resources. Providers can stop provision of the resources by notifying the Language Grid Operator. Language Resource Providers have no responsibility for any direct or indirect damage caused by use of Providers' resources.

4.3 Computation Resource Providers

The Language Grid Operator makes computation resources available for non-profit purposes. The Operator can, with Providers' consent, select or change the language resources that can be deployed on Providers' computation resources.

Computation Resource Providers have rights similar to Language Resource Providers. Computation Resource Providers can obtain the registered information of Language Service Users who access their computation resources. Language Grid Service Manager allows them to monitor statistics on how their computation resources are being used. Computation Resource Providers, however, should not obtain log data other than statistical information.

If necessary, Computation Resource Providers can suspend the use of their resources. Computation Resource Providers can stop their servers without reference to the Language Grid Operator. If some server fails or is halted, the Language Grid Operator should reconfigure the computation resources dynamically. Computation Resource Providers have no responsibility for any direct or indirect damage caused by use of Providers' resources.

4.4 Language Service Users

Language Service Users can use language resources and computation resources, but only for non-profit purposes. A Language Service User must obtain a valid *User ID* and *Password* to use the Language Grid. Language Service Users can allow participants in events or activities organized by Language Service Users to use the Language Grid. To avoid the fraudulent usage of language resources, however, Language Service Users should not allow the participants to discover the User ID and Password. For example, in the case of an NPO offering medical-interpreter services to

[10] It is prohibited to provide any language resource obtained from the Language Grid as paid services.

foreign patients to talk to hospital doctors, the NPO should not enter their User ID and Password in front of the patients. It is also strongly recommended that Language Service Users should change Passwords periodically, and whenever they use the Language Grid in public. Furthermore, in accessing language resources, Language Service Users must conform to the terms of use expressed in each Language Resource Provider's copyright notice and/or licensing policy.

Language Service Users cannot be anonymous so as to satisfy the requirements of the Language Resource Providers. Language Service Users must agree that their statistical information will be collected and offered to the providers of the language and computation resources.

4.5 Language Grid Operator

The role of the Language Grid Operator is to conclude agreements with Language Grid Users, and to operate language and computation resources in the manner described in the agreements. Since the agreements are solely for non-profit use, the Operator is not to be involved in any for-profit contracts.

The Operator will not store or obtain any unpublished personal information. Language Grid Users are asked to input personal information when registering with the Language Grid. This information will be published on the Language Grid Operator's Web site. Also, the Operator should not obtain any usage log data other than statistics on the use of resources. All other information obtained is made public on the Operator's Web site.

Finally, the Language Grid Operator has no responsibility for any direct or indirect damage caused by use of the language and/or computation resources.

5 Discussion

Below we address the questions expected to be raised by the proposed operation model.

The basic question is how the Operator can guarantee that the language resources will not be used for profit. The Operator and Users should conclude agreements to use language resources solely for non-profit, but this cannot guarantee the non-profit use of resources. In response, the Operator will conclude agreements only with trustworthy organizations and individuals. Risk of leaking the ID and password cannot also be ignored. To prevent this, the Operator should always monitor the Language Grid. If resource abuse is detected, the Operator halts access to the language resources by the offending party.

Another question is the possibility of a large number of individuals using the Language Grid as a free resource. According to the model, non-profit use of the Language Grid includes personal use. This is because individuals may provide language resources. Therefore, Language Grid Operator will conclude agreements only with trustworthy individuals.

Can the Language Grid be used in the research labs of companies? According to the model, profit organizations can use the Language Grid only for social contribution activities, and cannot use it in their research labs. This is because language resource providers sometimes charge the research labs of companies for their resources. We need to avoid conflicts between existing businesses and the non-profit operation model of the Language Grid, which is to explore new application areas of language resources.

The non-profit operation model might be too strict to encourage the propagation of the Language Grid: Language Service Users cannot be anonymous, companies cannot use the Language Grid, most individuals are excluded in practice, only trustworthy organizations can join, and so on. Indeed, the model cuts many potential applications of the Language Grid. The non-profit operation model is mainly intended to promote the research and development of technologies with field workers. In the future, we need a sustainable model for operation, most likely some combination of for-profit and non-profit models, so that companies are encouraged to join the Language Grid, and provide high quality services to anyone in the world, just as Internet providers do at present.

6 Conclusion

This paper proposes a non-profit operation model to coordinate four types of stakeholders: Language Resource Provider, Computation Resource Provider, Language Service User, and Language Grid Operator. The essence of the model is as follows.

- Language Grid Operator concludes agreements among stakeholders and monitors the Language Grid so as to maintain the agreements. The Op-

erator holds no private personal information and will not be involved in for-profit activities.
- Language Resource Providers and Computation Resource Providers can monitor the usage of their resources and control access to their resources.
- Language Service Users, which are often organizations like NPOs and NGOs, can allow participants of their events and activities to access the Language Grid.

The operation model is designed to coordinate stakeholders and to match their incentives. Trial operation of the proposed model started in December 2007 with around thirty organizations[11]. The Language Grid software has been developed by NICT, while its operation is done by Department of Social Informatics, Kyoto University. A user group called Language Grid Association[12] including NOPs, NGOs, universities, and research institutes has been formed to develop and share experiences of using the Language Grid in their activities. An analysis of the results will contribute to the increased accessibility and usability of language resources and so overcome language barriers worldwide.

References

We give thanks to the member of Language Grid Association for their continuous collaboration. This research is partially supported by Kyoto University Global COE Program: Informatics Education and Research Center for Knowledge-Circulating Society.

References

U Callmeier, A Eisele, U Schafer and M Siegel. 2004. The Deep Thought Core Architecture Framework. *Proceedings of LREC*, pp.1205-1208.

N Calzolari, A Zampolli, and A Lenci. 2002. Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. *CICLing*, pp. 264-279.

K Choukri. 2004. European Language Resources Association History and Recent Developments. *SCALLA Working Conference* KC 14/20.

C Fellbaum and P Vossen. 2007. Connecting the Universal to the Specific: Towards the Global Grid. *Intercultural Collaboration*, Lecture Notes in Computer Science 4568, pp. 1-16.

D Ferrucci and A Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, Vol. 10, pp. 327-348 Cambridge University Press.

Y Hayashi and T Ishida. 2006. A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons. *International Conference on Language Resources and Evaluation*, pp.1-6.

Y Hayashi, T Declerck, P Buitelaar, and M Monachini. 2008. Ontologies for a Global Language Infrastructure. The First International Conference on Global Interoperability for Language Resources.

T Ishida. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. *IEEE/IPSJ Symposium on Applications and the Internet*, pp. 96-100.

T Ishida, S R Fussell and P Vossen Eds. 2007. *Intercultural Collaboration*. Lecture Notes in Computer Science, 4568, Springer-Verlag.

R Khalaf, N Mukhi, S Weerawarana. 2003. Service-Oriented Composition in BPEL4WS. *Proceedings of the World Wide Web Conference*.

S Nomura, T Ishida, N Yamashita, M Yasuoka and K Funakoshi. 2003. Open Source Software Development with Your Mother Language: Intercultural Collaboration Experiment 2002. *International Conference on Human-Computer Interaction*, Vol. 4, pp. 1163-1167, 2003.

P Vossen. 1998. Introduction to EuroWordNet. *Computers and the Humanities*, Vol. 32, No. 2-3, pp. 73-89.

N Yamashita and T Ishida. 2006. Effects of Machine Translation on Collaborative Work. *International Conference on Computer Supported Cooperative Work*, pp. 515-523.

[11] http://www.langrid.org/operation/.
[12] http://www.langrid.org/association/.

SHARABLE TYPE SYSTEM DESIGN FOR TOOL INTER-OPERABILITY AND COMBINATORIAL COMPARISON

Yoshinobu Kano[1] **Ngan Nguyen**[1] **Rune Sætre**[1] **Keiichiro Fukamachi**[1]
Kazuhiro Yoshida[1] **Yusuke Miyao**[1] **Yoshimasa Tsuruoka**[3]
Sophia Ananiadou[2,3] **Jun'ichi Tsujii**[1,2,3]

[1]Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 Tokyo

[2]School of Computer Science, University of Manchester
PO Box 88, Sackville St, MANCHESTER M60 1QD, UK

[3]NaCTeM (National Centre for Text Mining), Manchester Interdisciplinary Biocentre,
University of Manchester, 131 Princess St, MANCHESTER M1 7DN, UK

{kano,nltngan,satre,keif,kyoshida,yusuke,tsujii}
@is.s.u-tokyo.ac.jp
{yoshimasa.tsuruoka,sophia.ananiadou}@manchester.ac.uk

Abstract

Nowadays, an increasing number of language resources including both corpora and tools for Text Mining (TM) and Natural Language Processing (NLP) are available. Because most of TM/NLP tasks are composite by nature, the interoperability between tools (including corpora) becomes of increasing importance to integrate the combinations of tools and to further select a combination that would yield better results for a specific task and data. Although the generic frameworks like UIMA (Unstructured Information Management Architecture) provide promising ways to solve this problem, the solution they provide is only partial; we also need sharable type systems to obtain interoperability between independently developed tools. In this paper, we propose a way to design sharable type systems, which would allow the users to integrate combinations of tools, and to compare the combinations. We show its feasibility by our automatic combinatorial comparison generator that was developed based on UIMA, with a protein-protein interaction (PPI) extraction system as an example.

1 Introduction

Recently, an increasing number of TM/NLP tools such as part-of-speech (POS) taggers (Tsuruoka et al., 2005), named entity recognizers (NERs)(Settles, 2005), syntactic parsers (Hara et al., 2007; Pyysalo et al., 2006), and relation or event extractors (ERs), have been developed. However, it is still very difficult to integrate independently developed tools into an aggregated application that achieves a specific task. The difficulties are caused not only by differences in programming platforms and different input/output data formats, but also by the lack of the higher level interoperability among modules developed by different groups.

UIMA, Unstructured Information Management Architecture (Lally et al., 2004), which was originally developed by IBM and recently became an open project in OASIS and Apache, provides a promising framework for tool integration. Although it has a set of useful functionalities,

UIMA only provides a generic framework, thus it requires a user community to develop their own platforms with a set of actual software modules. A few attempts have already been made to establish platforms, e.g. the CMU UIMA component repository[1], or GATE (Cunningham et al., 2002) with its UIMA interoperability layer.

However, simply wrapping existing modules to be UIMA compliant does not offer a complete solution for flexible tool integration. Users, which include both developers and end-users of TM/NLP systems, tend to be confused when choosing appropriate modules for their own tasks from a collection of a large number of tools.

Individual TM/NLP user groups have diverse interests. Depending on these interests, requirements for TM/NLP modules vary significantly (Ananiadou et al., 2006). An NER module developed for a specific user group usually cannot satisfy the needs of another group. Different groups may need different types of entities to be recognized. They may also need to process different types of text, such as scientific papers, reports, or medical records. Due to this diversity, significant effort is often required to combine modules that were developed independently for different user groups, even after they are wrapped for UIMA.

Furthermore, most of TM/NLP tasks are composite in nature, and can only be solved by combining several modules. Although the selection of modules affects the performance of the aggregated system, it is difficult to estimate how this selection affects the ultimate performance of the system. Users need to test a large number of combinations of tools in order to pick a most suitable combination for their specific task.

Although *types* and *type systems* are the only way to represent meanings in the UIMA framework, UIMA does not provide any specific *types*, with the exception of a few purely primitive *types*. In this paper, we propose a way to design sharable *type systems*. A sharable *type system* designed in this way can provide the interoperability between independently developed tools with less of a loss in information, thus allowing combinations of tools and comparisons of combinations.

[1] http://uima.lti.cs.cmu.edu/

We show how our automatic comparison generator works based on a *type system* designed in that way. Taking extraction of protein-protein interaction (PPI) as a typical example of a composite task, we illustrate how our platform helps users construct a system based on their own needs.

2 Motivation and Background

2.1 Goal Oriented Evaluation, Module Selection and Inter-operability

There are standard evaluation metrics for NLP modules such as precision, recall and F-value. For basic tasks like sentence splitting, POS tagging, and named-entity recognition, these metrics can be estimated using existing gold-standard test sets. However, accuracy measurements based on standard test sets are sometimes deceptive, since the accuracy may change significantly in practice, depending on the types of text and the actual tasks at hand.

For example, in the bioinformatics task of recognizing occurrences of entities of specific types (e.g. cell-lines, cell locations) in text when comprehensive lexicons for those entities are available, an NER for an open set of entities (e.g. proteins, metabolites, etc.) trained using a gold-standard training set may not be the best choice, even if it has the best performance on a standard test set. Moreover, systems which have similar levels of performance, according to standard metrics often behave differently in specific cases. Because these accuracy metrics do not take into account the importance of the different types of errors to any particular application, the practical utility of two systems with seemingly similar levels of accuracy may in fact differ significantly. To users and developers alike, a detailed examination of how systems perform (on the text they would like to process) is often more important than standard metrics and test sets. Naturally, far greater importance is placed in measuring the end-to-end performance of a composite system than in measuring the performance of individual components.

In reality, because the selection of modules usually affects the performance of the entire system, the careful selection of modules that are appropriate for a given task is crucial. This is the main reason for having a collection of

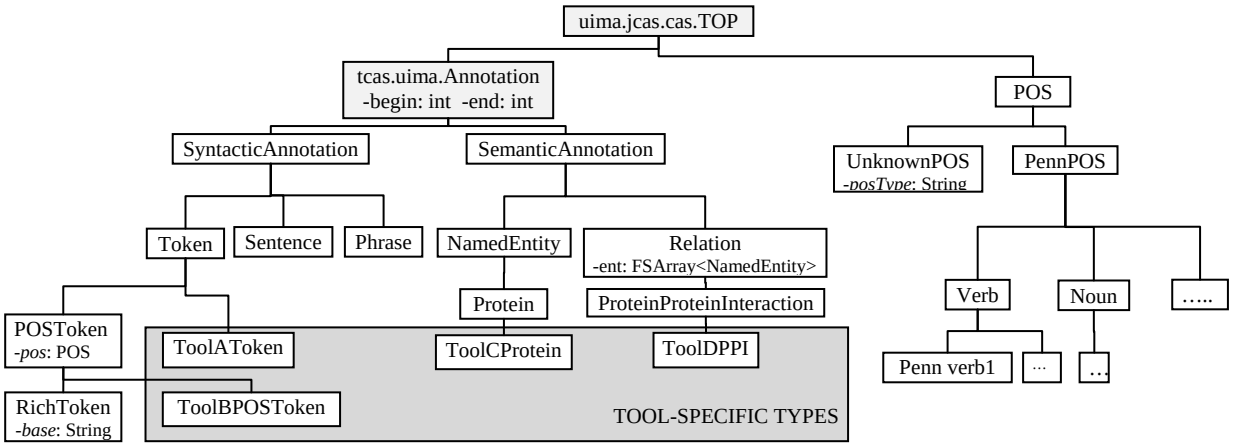


Figure 1. Part of our type system

interoperable modules. What we need is to show how the ultimate performance will be affected by the selection of different modules and what would be the best combination of modules in terms of the performance of the whole aggregated system for the task at hand.

Since the number of possible combinations of component modules is typically large, the system has to be able to enumerate and execute them semi-automatically. This requires a higher level of interoperability of individual modules than just wrapping them for UIMA.

2.2 UIMA

2.2.1 CAS and Type System

The UIMA framework uses the "stand-off annotation" style (Ferrucci et al., 2006). The raw text in a document is kept unchanged during analysis, and when the processing on the text is performed, the result is added as new stand-off annotations with references to their positions in the raw text. A *Common Analysis Structure* (*CAS*) maintains a set of these annotations, which in turn are objects by themselves. The annotation objects in a *CAS* belong to *types* that are defined separately in a hierarchical *type system*. The features of an *annotation*[2] object have values, which are typed as well.

2.2.2 Component and Capability

Each UIMA component has a *capability* property, which describes what *types* of objects the component may take as its input and what *types* of objects it produces as its output. For example, a named entity recognizer detects named entities in the text and outputs annotation objects of the *type* `NamedEntity`.

It is possible to deploy any UIMA component as a SOAP web service, so that we can combine a remote component on a web service with local component freely inside a UIMA-based system.

3 Sharable Type System for Combinatorial Comparison

Although UIMA provides a set of useful functionalities for an integration platform of TM/NLP tools, users still have to develop the actual platform by using these functionalities effectively. It is crucial how to define *types* and a *type system* in UIMA. In this section, we discuss formal characteristics of a sharable type system, which are needed for automatic combinations of components and comparison of combinations. Our discussion is based on the UIMA framework, but it could be applied to any framework that defines types and type systems.

3.1 Sharable Type System

When different groups develop different *type systems*, we should convert *types* between the different *type systems* to connect the components of the different *type systems*. Because it is cumbersome to make type converters for each pair of *type systems*, we need a *sharable type system*. Firstly, we discuss the characteristics of a *type system*, which can be shared by different groups.

[2] In the UIMA framework, `Annotation` is a base *type* which has *begin* and *end* offset values. In this paper we call any objects (any subtype of `TOP`) as *annotations*.

One extreme is to wrap existing programs without using explicit *types*, thus putting information to a single String field of a common generic *type*. Since the compatibility among modules is already automatically guaranteed, such a design decision would be easy to follow. However, it would not be appropriate if our aim is to attain the higher level of inter-operability: the represented meaning would not be unique; it cannot represent relations between meanings.

Another extreme is to enforce all modules developed by different groups to accept a unique *type system* defined by a platform. While this makes inter-operability readily attainable, it places too much of a burden on individual modules. In the worst case, we may have to re-program all the programs developed by other groups, which make this design decision impossible.

Our decision lies in the middle between these two extremes. That is, if necessary, we keep different *type systems* by individual groups as they are. However, we require that individual *type systems* have to be related through a common, sharable *type system*, which our platform defines. Such a shared *type system* can bridge modules with different *type systems*, though in bridging modules, we may lose some information during the translation process.

The most natural way to bridge local *type systems* through a sharable *type system* is to make local *types* subtypes of sharable *types* [3] . The sharable *type system* and local *type systems* can be considered to form a large single *type system* in this case; both of the local *type system* hierarchy and the sharable *type system* hierarchy are retained and accessible by users. Each *type system* (shared or local) is supposed to represent consistent concepts. The local *type system* can be developed and maintained independently.

In order to maintain information during the subtype bridging process, *type system* should be as hierarchical as possible. It is better to expand features as *types* rather than to use feature values, if the features are considered as a finite set of values. It would require multiple inheritances.

This bridging strategy with multiple inheritances requires local *type systems* to be hierarchical. If the local *type system* is not well-formed in its hierarchy, it may be difficult to bridge *type systems* in this way. We should prepare a type converter in such cases, though we cannot use the sharable and local *type system* information at the same.

Whether such a sharable *type system* can be defined easily or not is dependent on the nature of each problem. For example, a sharable *type system* for POS tags in English can be defined rather easily, since most of POS-related modules such as POS taggers (their output is a sequence of POSs), shallow parser (their input is a sequence of words with their POS assignments), etc. follow, more or less, the well established types defined by the Penn Treebank tag set for POS *types*.

Figure 1 shows a part of our shared *type system*. We deliberately define a highly organized *type* hierarchy, since the structure of the sharable *type system* directly influences the loss of information during the translation process.

3.2 General Combinatorial Comparison Generator and Type System

Secondly, we require the *sharable type system* to be used to automatically make possible combinations of the components. We illustrate these issues using the PPI *workflow* that we utilized in our experiments.

Figure 2 shows the *workflow* of our whole PPI system conceptually. If we can prepare two or more components for some type of the components in the *workflow* (e.g. two sentence detectors and three POS taggers), then we could make the combinations of these tools to form a multiplied number of *workflow* patterns (two sentence detectors x three POS taggers = 6 patterns). See Table 1 for the details of UIMA components used in our experiments.

We made a pattern expansion mechanism which generates possible *workflow* patterns automatically from a user-defined *comparable workflow*. A *comparable workflow* is a special *workflow* which explicitly specifies which set of UIMA components should be compared. Then, users just need to group comparable components (e.g.

[3] It requires multiple inheritances in the *type system*. Although the current *type system* of Apache UIMA implementation is a simple tree structure, multiple inheritances will be provided in the future releases because the UIMA *type system* is documented as ECore compatible, while ECore allows multiple inheritances.

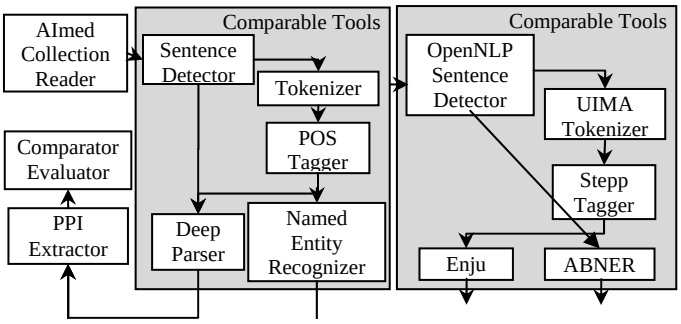


Figure 2. PPI system workflow (conceptual)

Figure 3. Basic example pattern

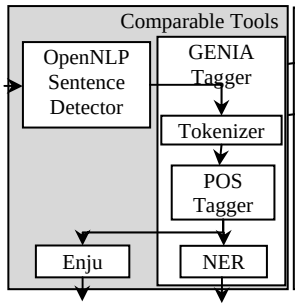


Figure 4. Complex tool example

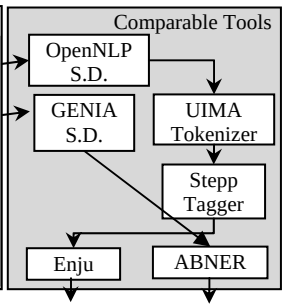


Figure 5. Branch flow pattern

ABNER[4] and MedT-NER as a comparable NER group) without making any modifications to the original UIMA components. This aggregation of comparable components is controlled by our *custom workflow controller*.

Types should be defined in a distinct and hierarchical manner to allow such an automatic expansion. For example, both tokenizers and POS taggers output tokens, but their roles are different when we assume a cascaded pipeline. We defined `Token` as a supertype and `POSToken` as a subtype of `Token`. Each component should have an individual *type* to clarify which component generated which instance, because each component may have a slightly different definition though their *types* tend to have a similar name. This is important because the *capabilities* are represented by these *types*, and the *capabilities* are only attributes that are machine readable. Such an individual *type* for each component is also needed to enable the automatic combinations without an infinite search loop.

In some cases, a single tool can play two or more roles (e.g. the GENIA Tagger performs tokenization, POS tagging, and NER; see Figure 4). It may be possible to decompose the original tool into single roles, but in most cases it is difficult and unnatural to decompose such a complex tool. We designed our comparator to detect the possible input combinations automatically by the *types* of previously generated *annotations*, and the input *capability* of each posterior component. As described earlier, a component should have appropriate *capabilities* with proper *types* in order to permit this detection.

When a component requires two or more input *types* (e.g. our PPI extractor requires outputs of a deep parser and a protein NER system), there could be a different set of components used in the prior flow (e.g. OpenNLP and GENIA sentence detectors in Figure 5). Thus, our comparator calculates such cases automatically.

The entire *type system* should not contain cyclic dependencies, i.e. it should be acyclic. A cyclic *type system* may make the automatic expansion indeterministic.

3.3 Comparable Type System

Finally, we should consider that the *type system* could be used to compare a similar sort of components. Any two *types*, which are considered to be comparable, should have a same ancestor *type*.

However, it is difficult to decide which *types* have specific relations; "definitions" or "meanings" of *types* are unclear in most cases. For example, it is really difficult to strictly define what `Sentence`, `Protein`, `Token`, etc. are. From the computational point of view, we could only observe *distributions* of instances of a *type*. A *distribution* is a set of *occurrences* of the instances, where an *occurrence* of an instance is characterized by its feature values. For example, in the TM/NLP field, one of the feature values are typically offset positions in the text. When the feature values of two instances are same, they can be considered as identical.

Distribution of a parent *type* should contain *distribution* of its child *type*, at least conceptually, because abstract *types* could be used to compare instances of all subtypes.

The experts of specific domains are required to complete an actual entire *type system* design. For

[4] In the example figures, ABNER requires `Sentence` to make the explanation clearer, though ABNER does not require it in actual usage.

Sentence	Token	POSToken	RichToken	Protein	Phrase	PPI
GENIA Tagger: Trained on the WSJ, GENIA and PennBioIE corpora (POS). Uses Maximum Entropy (Berger et al., 1996) classification, trained on JNLPBA (Kim et al., 2004) (NER). Trained on GENIA corpus (Sentence Splitter).						
Enju: HPSG parser with predicate argument structures (PAS) as well as phrase structures. Although trained with Penn Treebank, it can compute accurate analyses of biomedical texts owing to its method for domain adaptation (Hara et al., 2005).						
STePP Tagger: Based on probabilistic models, tuned to biomedical text trained by WSJ, GENIA (Kim et al., 2003) and PennBioIE corpora.						
MedT-NER: Statistical recognizer trained on the JNLPBA data.						
ABNER: From the University of Wisconsin (Settles, 2005), wrapped by the Center for Computational Pharmacology at the University of Colorado.						
Akane++: A new version of the AKANE system (Yakushiji, 2006), trained with SVMlight-TK (Bunescu et al., 2006; Joachims, 1999; Moschitti, 2006) and the AImed Corpus.						
Annotation Comparator and Evaluator: Compares annotations using the type system hierarchy to decide which annotations can be compared; generates statistical results and visualization.						
UIMA Examples: Provided in the Apache UIMA example. Sentence Splitter and Tokenizer.						
OpenNLP Tools: Part of the OpenNLP project (http://opennlp.sourceforge.net/), from the Apache UIMA examples.						
AImed Corpus: 225 Medline abstracts with proteins and protein-protein interactions annotated (Bunescu et al., 2006).						

Legend: Input type(s) required for that tool Input type(s) required optionally Output type(s)

Table 1. List of UIMA-compliant tools that we used in the experiment.

instance, there are enormous numbers of possible abstract *types*, of which only experts can tell which *types* are really meaningful. However, the discussion of this section provides a rough but fundamentally important direction for the *type system* design.

4 Experiments and Results

We have performed experiments using our PPI extraction system as an example (Kano et al., 2008). The PPI system (Figure 2) is similar to our BioCreative PPI system (Sætre et al., 2007). It differs in that we have broken up the original system into seven different components.

As summarized in Table 1, we have several comparable components in addition to the original PPI system, and AImed as gold standard data. In this case, possible combination *workflow* patterns are 36 for `PosToken`, 589 for `ProteinProteinInteraction`, etc.

Table 2 and Figure 6 show a part of the comparison result screenshots between these patterns on 20 articles from the AImed corpus. In Table 2, *labeled* scores represent complete matches of every feature of *annotations*, while unlabeled scores ignore primitive fields excluding offsets (e.g. compare offsets but ignore protein IDs). Table 3 shows a part of PPI extraction results from which we can discern which combination of tools generate the best result.

.POSToken .POS	.Sentence	.SimpleToken	gold	OpenNLP Genia OpenNLP 3105	Stepp Genia OpenNLP 1693
Stepp	Genia	UIMA	1661	40/75 (40/75)	62/63 (62/6
Stepp	AIMED	UIMA	1661	40/75 (40/75)	62/63 (62/6
GENIA	Genia	UIMA	1661	39/73 (39/73)	53/54 (53/5
GENIA	AIMED	UIMA	1661	39/73 (39/73)	53/54 (53/5
Stepp	AIMED	GENIA	1690	31/57 (31/57)	63/63 (63/6
Stepp	AIMED	OpenNLP	1692	31/57 (31/57)	63/63 (63/6

Table 2. A Screenshot of a combinatorial comparison for *type* `Protein`. Values are precision/recall in "labeled (unlabeled)" pairs, and total numbers of instances are shown.

When neither of the compared results include the gold standard data (AImed in this case), the comparison results show a *similarity* of the tools for this specific task and data, rather than an evaluation. Even if we lack an annotated corpus, it is possible to run tools and compare results in order to understand the characteristics of tools depending on the corpus and the tool combinations.

5 Conclusion and Future Work

Although UIMA provides a general framework with much functionality, we still need to fill in the gaps between what is already provided and what the users need for their specific tasks. NLP tasks typically consist of many components, and it is necessary to show which set of tools are most suitable for each specific task and each specific datum. In this paper we provided an answer to this problem using the extraction of protein-protein interaction as an example task.

The *type system* design is one of the most critical issues on the interoperability, which the UIMA framework does not provide. We proposed a way to design a sharable *type system* as a bridge between locally defined *type systems*. We also discussed a *type system* design, which allows for the automatic combinations of tools and comparisons of these combinations.

With any set of UIMA compliant components which have *types* designed in the way described in this paper, our general combinatorial comparator generates possible combinations of tools for a specific *workflow* and compares/evaluates the results. We are preparing to make a portion of the components and services described in this paper available publicly (http://www-tsujii.is.s.u-tokyo.ac.jp/uima/ and http://u-compare.org/).

The final system shows the combination of components that has the best score, but it also generates comparative results. This helps the users grasp the characteristics and differences between the tools which cannot be easily observed just by the widely used F-score evaluations.

Future directions for this work includes combining the output of several modules of the same kind (such as NERs) to obtain better results, collecting other tools developed by other groups using bridging *type systems*, making the coverage of the sharable *type system* to be broadened, and making grid computing available with UIMA *workflows* to increase the entire performance without modifying original UIMA components.

Features	Prec	Recall	F1
DEP	67.6	26.3	37.1
WORDS	55.7	29.2	37.8
PAS (Enju)	72.0	28.7	41.0
DEP+WORDS	59.9	39.3	46.9
PAS+DEP	68.9	37.8	48.6
PAS+WORDS	61.3	40.7	48.6
ALL	64.3	44.1	52.0
ALL (pairwise)	78.1	62.7	69.5

Table 3. PPI Evaluated on AImed, with 5631 protein pairs. (1068 true interactions). DEP means our dependency parser. Values are percentages from 10-fold cross-validation on abstracts. “pairwise” is the widely used 10-fold cross-validation on protein pairs.

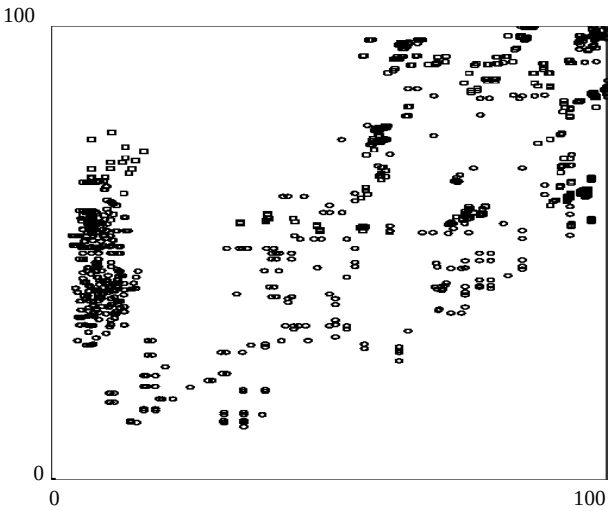


Figure 6. NER comparison distribution of precisions (x-axis) and recalls (y-axis).

Acknowledgments

We wish to thank Dr. Lawrence Hunter’s text mining group at the Center for Computational Pharmacology for discussing with us and making their tools available for this research. This work was partially supported by NaCTeM (the UK National Centre for Text Mining), Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan). NaCTeM is jointly funded by JISC/BBSRC/EPSRC.

References

Sophia Ananiadou, Douglas B. Kell, and Jun'ichi Tsujii. 2006. *Text mining and its potential applications in systems biology*. Trends in Biotechnology, 24(12), 571-579.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. *A maximum entropy approach to natural language processing*. Comput. Linguist., 22(1), 39-71.

Razvan Bunescu, and Raymond Mooney. 2006. *Subsequence Kernels for Relation Extraction*. In Y. Weiss, B. Scholkopf and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (171--178). Cambridge, MA: MIT Press.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. *GATE: A framework and graphical development environment for robust NLP tools and applications.* In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.

David Ferrucci, Adam Lally, Daniel Gruhl, Edward Epstein, Marshall Schor, J. William Murdock, et al. 2006. *Towards an Interoperability Standard for Text and Multi-Modal Analytics* (No. RC24124): IBM Research Report.

Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. 2005. *Adapting a probabilistic disambiguation model of an HPSG parser to a new domain*. In *Natural Language Processing - Ijcnlp 2005, Proceedings* (Vol. 3651, 199-210). Berlin: Springer-Verlag Berlin.

Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. 2007, June. *Evaluating Impact of Re-training a Lexical Disambiguation Model on Domain Adaptation of an HPSG Parser*. In Proceedings of the IWPT 2007, Prague, Czech Republic.

Thorsten Joachims. 1999. *Making large-Scale SVM Learning Practical*. In Schölkopf B., C. Burges and A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning* (169-184). Cambridge, MA: MIT Press.

Yoshinobu Kano, Ngan Nguyen, Rune Sætre, Kazuhiro Yoshida, Yusuke Miyao, Yoshimasa Tsuruoka, et al. 2008, January. *Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example.* In Proceedings of the Pacific Symposium on Biocomputing, Hawaii, USA.

Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi, and Jun'ichi Tsujii. 2003. *GENIA corpus - a semantically annotated corpus for bio-textmining*. Bioinformatics, 19(suppl. 1), i180-i182.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, and Yuka Tateisi. 2004. *Introduction to the Bio-Entity Recognition Task at JNLPBA.* In Proceedings of the Proceedings of the International Workshop on Natural Language Processing, Geneva, Switzerland.

Adam Lally, and David Ferrucci. 2004. *Building an Example Application with the Unstructured Information Management Architecture*. IBM Systems Journal, 43(3), 455-475.

Alessandro Moschitti. 2006. *Making Tree Kernels Practical for Natural Language Learning.* In Proceedings of the Eleventh International Conference on European Association for Computational Linguistics, Trento, Italy.

S. Pyysalo, T. Salakoski, S. Aubin, and A. Nazarenko. 2006. *Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches*. BMC Bioinformatics, 7, 9.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007, April. *AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask.* In Proceedings of the Second BioCreative Challenge Evaluation Workshop.

Burr Settles. 2005. *ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text.* Bioinformatics, 21(14), 3191-3192.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, and Tomoko Ohta. 2005. *Developing a Robust Part-of-Speech Tagger for Biomedical Text.* In Proceedings of the Advances in Informatics - 10th Panhellenic Conference on Informatics, Volos, Greece.

Akane Yakushiji. 2006. *Relation Information Extraction Using Deep Syntactic Analysis.* Ph.D Thesis, University of Tokyo.

Encoding Hierarchical Bilingual Texts of Hong Kong Laws with XCES

Chunyu Kit and **Hio Tong Chan** and **Xiaoyue Liu**
Department of Chinese, Translation and Linguistics
City University of Hong Kong, Tat Chee Ave., Kowloon, Hong Kong
E-mail: {ctckit, cthilda, xyliu0}@cityu.edu.hk

Abstract

This paper presents our recent work on encoding the hierarchical English-Chinese bilingual BLIS corpus of HK laws following XCES. Our work is aimed not only at facilitating data capture for language engineering such as machine translation but also at laying down a foundation for a suitable presentation of legislative texts for various purposes (e.g., rendering the loose-leaf edition of HK laws on the Web and supporting online bilingual legislative drafting). Relevant corpus encoding standards and the text hierarchy of HK laws are reviewed, and the technical details for encoding linguistic annotation and text alignment at all hierarchical levels are presented.

1 Introduction

Corpora are essential resources for both linguistic research and language engineering. The past decades have observed a significant advance in different aspects of corpus development. From Brown corpus (Francis and Kucera, 1979) and PTB[1] (Marcus, Santorini, and Marcinkiewicz, 1993) to BNC[2] (Aston and Burnard, 1998) and ANC[3] (Reppen and Ide, 2004; Ide and Suderman, 2006) and then to CGW[4] (Ma and Huang, 2006), we can see that the scale of corpus size has increased at least a thousand times, from one million words to a hundred million and then to over a billion. Besides monolingual corpora, many bilingual and multilingual parallel corpora have also been developed for various applications, e.g., Canadian Hansard[5] (Brown, Lai, and Mercer, 1991; Gale and Church, 1991) and HK Hansard[6] (Wu, 1994), Chinese-English Parallel Corpus (Sun et al., 2002), and the JRC-Acquis Multilingual Parallel Corpus[7] (Ralf et al., 2006).

All corpora have to be encoded in a suitable format for data exchange and/or other purposes of research and engineering. To encode more complicated information in a corpus, we need a more sophisticated annotation scheme. Along this trend, corpus annotation has evolved from the elemental formats, such as the simple POS tag attachment and syntactic tree bracketing as used in PTB, to various sophisticated markup languages such as SGML and XML, as illustrated in BNC and ANC respectively. Consequently, CES and its XML version XCES are developed[8] (Ide, 1998; Ide, Bonhomme, and Romary, 2000).

In this paper, we present our recent work on encoding the BLIS corpus, the most comprehensive and authoritative collection of English-Chinese bilingual texts of HK laws, using and also extend-

[1] The Penn Treebank Corpus, 3rd ed., at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42.

[2] British National Corpus, at http://www.natcorp.ox.ac.uk/.

[3] American National Corpus, at http://www.americannationalcorpus.org/.

[4] Chinese GigaWord Corpus, 2nd ed., at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T14.

[5] Hansard French/English, at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20.

[6] Hong Kong Hansards Parallel Text, at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId= LDC2000T50.

[7] http://langtech.jrc.it/JRC-Acquis.html.

[8] Corpus Encoding Standard, at http://www.cs.vassar.edu/CES/; XCES: Corpus Encoding Standard for XML, at http://www.cs.vassar.edu/XCES/.

ing the XCES schemas. The purpose of our work is multifold. It is aimed not only at conducting linguistic annotation of parallel texts as resources for language engineering tasks such as machine translation but also at laying down a sound foundation for bilingual legal text processing and document management. In particular, we are interested in rendering the conventional printed version of HK laws for Web access and in developing facilities to support bilingual legislative drafting, with the aid of the standardized XML markup for corpus encoding.

The rest of the paper is organized as follows. The next section gives an overview of relevant corpus encoding standards. The text hierarchy of HK laws is then introduced in Section 3. Section 4 illustrates how XCES schemas are used and extended to encode the BLIS corpus, including the annotation for its primary data, morpho-syntactic description, and text alignment. A Web rendering of the print format of HK laws is also illustrated based on this annotation. The paper is then concluded in Section 5.

2 Corpus Encoding Standards

BNC and ANC are two best-known large-scale English corpora. They are comparable in several ways, including corpus size, both of over 100 million words, linguistic information, and the inclusion of both written and transcribed spoken data. The greatest difference between them, however, is their annotation formats: BNC was encoded with SGML (ISO 8879, 1986; Burnard, 1999), whereas ANC with XCES, the XML version of CES. Both of them comply with the TEI Guidelines[9] (Sperberg-McQueen and Burnard, 1994; Sperberg-McQueen and Burnard, 2007). Noticeably, however, the latest version of BNC has been turned into XML format since early 2007,[10] reflecting the current trend of corpus encoding towards using XML in the field.

2.1 TEI and CES

TEI was established in 1987 to provide universal text encoding standard for encoding language resources and has been widely accepted by researchers and enterprises to represent linguistic texts (Ide and Vronis, 1995). It is an application of SGML/XML with a set of DTD tagsets to specify the allowable tags and their attributes within a constructed text hierarchy for encoding text creation and data capture. Over 130 projects, including BNC and GENIA (Erjavec et al., 2003), have adopted its guidelines for corpus encoding. It has been through several versions, among which the latest one TEI P5 was released in 2005 (Sperberg-McQueen and Burnard, 2007).

Nevertheless, TEI is designed to cover a huge range of applications. In order to have it properly and specifically tailored for corpus encoding, CES is developed as a part of EU EAGLES[11] to provide an SGML DTD as a parameterized realization of TEI for this particular purpose.

2.2 CES and XCES

XCES results from an XML conversion of CES by converting its DTD into a set of XML schemas. Conforming to the same guidelines, they are implemented with difference markup languages. CES is an application of SGML via a parameterized DTD of TEI, whereas XCES is an application of XML through a number of schemas.

XML has been the most popular and, actually, the *de facto* standard markup language for linguistic data representation and exchange within the NLP community. It has been successfully applied to various kinds of annotation task (Muller, 2006) and also has received nice technical support from the industry, e.g., Sun Java's JAXP,[12] beyond the essential tools such as XSLT[13] to transform XML data into other formats (e.g., HTML and plain text). There have been many corpora encoded using XML, including ANC and CroCo[14] (Hansen-Schirra, Neumann, and Vela, 2006).

3 BLIS Hierarchy

The current version of the BLIS corpus to be annotated is acquired from the Bilingual Laws Information System[15] (Kit et al., 2004; Kit et al., 2005). It

[9] The TEI Guidelines for Electronic Text Encoding and Interchange of the Text Encoding Initiative, http://www.tei-c.org/.

[10] BNC XML Edition, at http://www.natcorp.ox.ac.uk/XMLedition/.

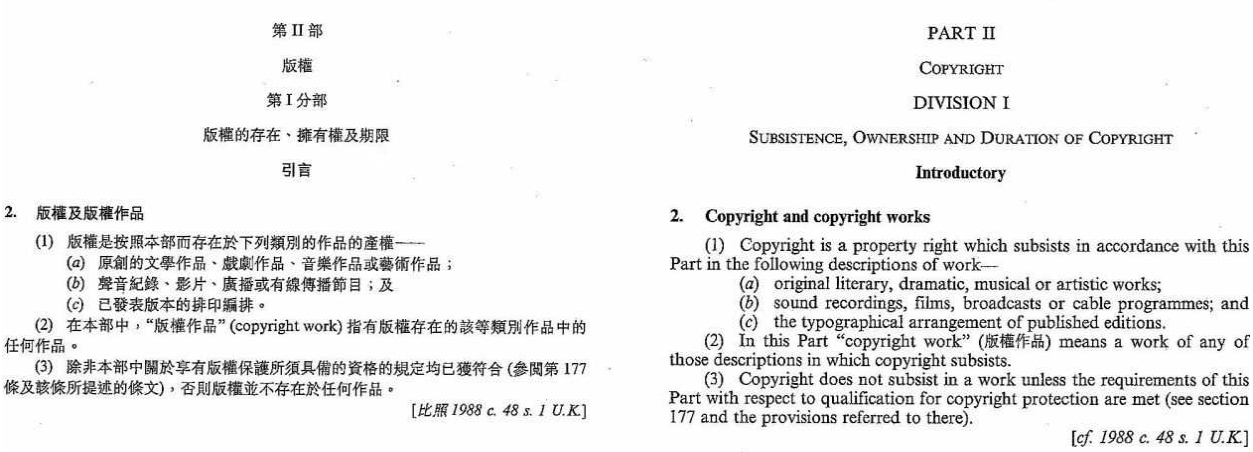
[11] The Expert Advisory Group on Language Engineering Standards, at http://www.ilc.cnr.it/EAGLES/home.html.

[12] Java API for XML Processing, at http://java.sun.com/xml/.

[13] EXtensible Style Language Transformation

[14] http://fr46.uni-saarland.de/croco/

[15] A searchable database of Hong Kong legislation freely accessible from http://www.legislation.gov.hk/.



第 II 部

版權

第 I 分部

版權的存在、擁有權及期限

引言

2. 版權及版權作品

(1) 版權是按照本部而存在於下列類別的作品的產權——

(*a*) 原創的文學作品、戲劇作品、音樂作品或藝術作品；

(*b*) 聲音紀錄、影片、廣播或有線傳播節目；及

(*c*) 已發表版本的排印編排。

(2) 在本部中，"版權作品" (copyright work) 指有版權存在的該等類別作品中的任何作品。

(3) 除非本部中關於享有版權保護所須具備的資格的規定均已獲符合 (參閱第 177 條及該條所提述的條文)，否則版權並不存在於任何作品。

[*比照 1988 c. 48 s. 1 U.K.*]

PART II

COPYRIGHT

DIVISION I

SUBSISTENCE, OWNERSHIP AND DURATION OF COPYRIGHT

Introductory

2. **Copyright and copyright works**

(1) Copyright is a property right which subsists in accordance with this Part in the following descriptions of work—

(*a*) original literary, dramatic, musical or artistic works;

(*b*) sound recordings, films, broadcasts or cable programmes; and

(*c*) the typographical arrangement of published editions.

(2) In this Part "copyright work" (版權作品) means a work of any of those descriptions in which copyright subsists.

(3) Copyright does not subsist in a work unless the requirements of this Part with respect to qualification for copyright protection are met (see section 177 and the provisions referred to there).

[*cf. 1988 c. 48 s. 1 U.K.*]

Figure 1: Illustration of the Loose-Leaf Edition of HK laws

is a complete collection of the status laws of HK, including all ordinances and subsidiary legislation of HKSAR and other relevant instruments. The entire corpus comprises 503 chapters, each for an individual ordinance, in a total of about 10 million English words and 18 million Chinese characters. Both languages are equally authentic by HK laws. All existing ordinances before the handover in 1997 were translated into Chinese by professionals and the later ones were legislated via bilingual laws drafting.

A unique numbering system is employed to represent the text hierarchy of HK laws in the following way, where an assigned number may be followed by an optional uppercase letter.

- Chapter, identified by an Arabic numeral, such as CAP 1, CAP 2, and CAP 3A.
- Part / Division / Subdivision / Subsubdivision, all optional, identified by an uppercase Roman numeral, such as PART I, DIV II.
- Section / Regulation / Rule / Bylaw / Article, identified by an Arabic numeral, such as 1, 2, 3. They all correspond to the same counterpart 条 in Chinese.
- Subsection, identified by an Arabic numeral in parentheses, such as (1), (2), (1A), and (2B).
- Paragraph, identified by a lowercase letter in parentheses, such as (a), (b).
- Subparagraph, identified by a lowercase Roman numeral in parentheses, such as (i), (ii).
- Subsubparagraph, identified by a uppercase letter in parentheses, such as (A), (B).
- Other optional items include Schedule, Order, and Appendix.

Both the Chinese and English versions of an ordinance observe exactly the same text hierarchy: chapter (章), part (部), division (分部), section (条), subsection (款), paragraph (段), subparagraph (节), and subsubparagraph (分节). Note that a text entity at the subsection level or below may contain only a sentence, a fragment of a sentence, a phrase or even a single word. Excerpts from the Authorized Loose-Leaf Edition of HK Laws published by the Department of Justice, HKSAR, are illustrated in Figure 1.

4 Encoding BLIS with XCES

The BLIS corpus was hierarchically aligned down to the subsubparagraph level by utilizing the above numbering system and encoded, accordingly, into a self-contained XML format (Kit et al., 2005). Now it is to be aligned down to the word level and encoded with the XCES schemas.[16] In particular, our work relies on three of them, namely, xcesDoc.xsd for primary data, xcesAna.xsd for tokenization and grammatical description, and xcesAlign.xsd for text alignment at the sentence and token levels.

We note, however, that the restrictive small tag set defined in xcesDoc.xsd only allows a circuitous way to encode the BLIS hierarchy. The readability

[16]The eight schemas provided by XCES are xcesDoc.xsd, xcesAna.xsd, xcesAlign.xsd, xcesWord.xsd, xcesSpoken.xsd, xcesHeader.xsd, xcesGlobal.xsd, and xcesLinks.xsd, available from at http://www.cs.vassar.edu/XCES/schema/.

of such encoding is unmanageable although the primary data could be completely encoded. The main cause for this is that XCES focuses more on encoding basic linguistic objects (like sentences and words) for the purpose of data capture rather than representing document structure and text hierarchy. To render a complicated text hierarchy as the one of the BLIS corpus, it is necessary to define our own schema with the aid of XCES.

4.1 Primary Data

Following the XCES recommendations, our first step towards the BLIS corpus encoding is to annotate its primary data, namely, its raw texts at each of its hierarchical level. Figure 2 illustrates a possible way of primary data encoding for a section of BLIS text, where the text structures at different levels need to be encoded with nested (text division) elements. Each may contain a number of subordinate elements such as <p> (paragraph), <list> (list) and <s> (sentence) for its subordinate hierarchical levels. Each of them carries a unique identifier as the value for its *id* attribute. All these *id* values form a referential system for alignment.

Note that there is a subtle but critical difference between a normal discourse paragraph and the paragraph in the BLIS hierarchy: one of the former, or even a long sentence, may contain several of the latter. This is why we have to use <s> to encode BLIS paragraph, subparagraph and subsubparagraph, and distinguish them from each other with the aid of the *type* attribute. This attribute needs to take on various values to specify the types of text structure. Working this way, however, we rely solely on one attribute to distinguish several text structures encoded by the same tag, inevitably resulting in a very tedious and circuitous encoding for the BLIS hierarchy, as shown in Figure 2.

One of our attempts is to encode BLIS primary data at all hierarchical levels and also at sentence boundaries following XCES. Nevertheless, the repeated use of each of the restrictive elements of XCES in such a heavy way does not seem to be an efficient path to rendering the complicated hierarchy of BLIS. The same drawback can also be found in others' works on legal document encoding using CES or XCES, e.g., Italian and German bilingual legal documents (Gamper, 2000) and French, German,

```
<div type="section" id="C528.S2">
 <head id="C528.S2.h">
  2.  Copyright and copyright works
 </head>
 <div type="remarks" id="C528.S2.re>
   <p id="C528.S2.re.p1>
     <s id="C528.S2.re.p1.s1></s>
   </p>
 </div>
 <div type="subsection" id="C528.S2.SS1>
  <p id="C528.S2.SS1.p1">
   <s id="C528.S2.SS1.p1.s1">(1) Copyright...
    <list>
     <item>
      <s type="paragraph" id="C528.S2.SS1.p1.s1.a">
       (a) original literary, dramatic, musical...
      </s>
     </item>
     <item>
      <s type="paragraph" id="C528.S2.SS1.p1.s1.b">
       (b) sound recordings, films, broadcasts...
      </s>
     </item>
     <item>
      <s type="paragraph" id="C528.S2.SS1.p1.s1.c">
       (c) the typographical arrangement of...
      </s>
     </item>
    </list>
   </s>
  </p>
 </div>
 <div type="subsection" id="C528.S2.SS2>
  <p id="C528.S2.SS2.p2">
   <s id="C528.S2.SS2.p2.s2">
    (2) In this Part "copyright work"...
   </s>
  </p>
 </div>
 <div type="subsection" id="C528.S2.SS3>
  <p id="C528.S2.SS3.p3">
   <s id="C528.S2.SS3.p3.s3">
    (3) Copyright does subsist in a work...
   </s>
  </p>
 </div>
 <note place="foot"></note>
</div>
```

Figure 2: Primary data annotation using XCES

Italian and Slovenian multilingual legal documents (Lyding et al., 2006). To remedy the deficiency of this kind and also enhance the accessibility and readability of the encoding for humans, we opt to extend XCES schemas by designing a set of intuitively readable tags on top of those available from XCES.

4.2 An Extended Schema for BLIS

A new schema is developed to provide a set of intuitively readable tags for all text structures in the BLIS hierarchy. It imports the whole xcesDoc.xsd schema for primary data annotation. A comparable situation is observed in the development of CES to extend the TEI guidelines into a restrictive encoding standard for linguistic corpora.

```
<section id="C528.S2">
 <head id="C528.S2.h">
  2.  Copyright and copyright works
 </head>
 <remark id="C528.S2.re"></remark>
  <subsection id="C528.S2.SS1">
   <p id="C528.S2.SS1.p1">
    <s id="C528.S2.SS1.p1.s1">(1) Copyright is...</s>
    <paragraph id="C528.S2.SS1.p1.s1.P1">
     <s id="C528.S2.SS1.p1.s1.P1.a">
      (a) original literary, dramatic, musical...
     </s>
    </paragraph>
    <paragraph id="C528.S2.SS1.p1.s1.P2">
     <s id="C528.S2.SS1.p1.s1.P2.b">
      (b) sound recordings, films, broadcasts...
     </s>
    </paragraph>
    <paragraph id="C528.S2.SS1.p1.s1.P3">
     <s id="C528.S2.SS1.p1.s1.P3.c">
      (c) the typographical arrangement of...
     </s>
    </paragraph>
   </p>
  </subsection>
  <subsection id="C528.S2.SS2">
   <s id="C528.S2.SS2.s2">(2) In this Part...</s>
  </subsection>
  <subsection id="C528.S2.SS3">
   <s id="C528.S2.SS3.s3">(3) Copyright does...</s>
  </subsection>
 <note place="foot"></note>
</section>
```

Figure 3: Primary data annotation with new schema

```
<xs:schema xmlns:xces="http://www.xces.org/schema/2003" ...
 <xs:include schemaLocation="xcesDoc.xsd" />
  <xs:complexType name="sectionType">
   <xs:annotation>
    <xs:documentation>
     The content of Section in BLIS. ...
    </xs:documentation>
   </xs:annotation>
   <xs:complexContent mixed="false">
    <xs:extension base="xces:class.text">
     <xs:sequence minOccurs="0" maxOccurs="unbounded">
      <xs:element name="head" type="headType" />
      <xs:element name="remarks" type="remarksTYpe" />
      <xs:choice minOccurs="0" maxOccurs="unbounded">
       <xs:element name="subSection" type="xces:subSectionType"/>
       <xs:element name="p" type="xces:subpar.seq" />
      </xs:choice>
     </xs:sequence>
     <xs:attribute name="type" type="xs:string" />
    </xs:extension>
   </xs:complexContent>
  </xs:complexType>
  <xs:complexType name="subsectionType">
   <xs:annotation>
    <xs:documentation>
     The content of subsection.  ...
    </xs:documentation>
   </xs:annotation>
   <xs:complexContent mixed="false">
    <xs:extension base="xces:class.text">
     <xs:sequence minOccurs="0" maxOccurs="unbounded">
      <xs:element name="p" type="xces:subpar.seq"/>
      <xs:element name="s" type="xces:subpar.seq"/>
     </xs:sequence>
    </xs:extension>
   </xs:complexContent>
  </xs:complexType>
...
</xs:schema>
```

Figure 4: A fragment of the new schema

The attribute *type* has a crucial role to play in the encoding using the xcesDoc.xsd schema. It specifies the type of an XCES element, e.g., , in correspondence to the text structure that it is applied to describe. The new schema is aimed at rectifying such overuse by providing a set of new elements such as <section>, <subsection>, <paragraph>, <subparagraph> and <subsubparagraph> to present the text hierarchy information in an intuitive way. As a core structure in the BLIS hierarchy, the Section and all its hierarchical equivalents such as Regulation, Rule, Bylaw and Article, and their Chinese counterpart 条, are all encoded as <section>. A <section> comprises a <head>, a <remark> and a <note>, to specify its referential title, remark and note respectively, in addition to one or more than one instance of <subsection> or <p>, exclusively, to present its content.

The <p> is retained for encoding a normal discourse paragraph and <paragraph> is specifically designed for a BLIS paragraph, which may comprise only a phrase or even a single word, as illustrated in Figure 1. The element <s> is retained for representing a sentence within <p> or <subsection>. It has a unique *id* value as the specifier for a sentence, which is to be used for sentence alignment. It may contain a few instances of <paragraph>.

Figure 3 illustrates the primary data encoding for the BLIS corpus using the new schema, a fragment of which is depicted in Figure 4. Comparing to the annotation using XCES directly, the current version shows a number of preferable features, including its naturalness, conciseness, readability and, especially, its nice correspondence to the BLIS hierarchy.

4.3 Linguistic Annotation

The basic task of linguistic annotation is to present linguistic information within the primary data. We focus on tokenization and POS tagging, and also on text alignment at both the sentence and token levels. The annotation outputs are stored in separate XML files with links to the correspondent primary data through the W3C XPointer.[17] This treatment follows the "stand-off" annotation strategy recommended by XCES.

[17] XML Pointer Language, at http://www.w3.org/TR/xptr-framework/, allows a hyperlink to point to a specific part in an XML document.

```
<chunkList xml:base="00528_s00002.sent.e.xml">
 <chunk type="sentence" xml:base="#C528.S2.SS1.s1">
  <tok id="C528.S2.SS1.s1w1" xlink:href="#xpointer
   (string-range(‘’,0,3))"/>
    <msd>CD</msd>
    <base>(1)</base>
  <tok id="C528.S2.SS1.s1w2" xlink:href="#xpointer
   (string-range(‘’, 4,9))"/>
    <msd>NN</msd>
    <base>copyright</base>
  <tok id="C528.S2.SS1.s1w3" xlink:href="#xpointer
   (string-range(‘’, 14,2))"/>
    <msd>VBZ</msd>
    <base>be</base>
  <tok id="C528.S2.SS1.s1w4" xlink:href="#xpointer
   (string-range(‘’, 17,1))"/>
    <msd>DT</msd>
    <base>a</base>
  <tok id="C528.S2.SS1.s1w5" xlink:href="#xpointer
   (string-range(‘’, 19,8))"/>
    <msd>NN</msd>
    <base>property</base>
...
 </chunk>
</chunkList>
```

Figure 5: Token annotation

```
<translations>
 <translation trans.loc="00528_s00002.sent.e.xml"
  xml:lang="en" n="1">
 <translation trans.loc="00528_s00002.sent.c.xml"
  xml:lang="big5" n="2">
</translations>
...
<linkList>
 <LinkGrp targType="s">
  <link>
   <align xlink:href="#C528.S2.SS1.s1"/>
   <align xlink:href="#C528.S2.SS1.s1"/>
  </link>
  <link>
   <align xlink:href="#C528.S2.SS1.s1.P1.a"/>
   <align xlink:href="#C528.S2.SS1.s1.P1.a"/>
  </link>
...
 </LinkGrp>
</linkList>
```

Figure 6: Sentence alignment

The annotation for tokenization and POS tagging using the xcesAna.xsd schema is illustrated in Figure 5. A token is encoded by <tok>, within which its POS and lemma are encoded in <msd> (morpho-syntactic description) and <base> respectively. A unique *id* is assigned to each <tok>, which is to be used for alignment. The XPointer in the attribute *xlink* is to associate the primary data and the stand-off annotation. A referential sentence for identifying token boundaries is provided through the attribute *xml:base* in a <chunk>, which is used to encode a sentence in our annotation for tokenization.

The text alignment at both the sentence and token levels is encoded using the xcesAlign.xsd schema. In fact, the encoding becomes a routine practice

```
<translations>
 <translation trans.loc="00528_s00002.tok.e.xml"
   xml:lang="en" n="1">
 <translation trans.loc="00528_s00002.tok.c.xml"
  xml:lang="big5" n="2">
</translations>
...
<linkList id="C528.S2.SS1.s1">
 <LinkGrp targType="tok">
  <link>
   <align xlink:href="#C528.S2.SS1.s1w1"/>
   <align xlink:href="#C528.S2.SS1.s1w1"/>
  </link>
  <link>
   <align xlink:href="#C528.S2.SS1.s1w2"/>
   <align xlink:href="#C528.S2.SS1.s1w2"/>
  </link>
  <link>
   <align xlink:href="#C528.S2.SS1.s1w3"/>
   <align xlink:href="#xces:undefined"/>
  </link>
  <link>
   <align xlink:href="#C528.S2.SS1.s1w4"/>
   <align xlink:href="#C528.S2.SS1.s1w13"/>
  </link>
...
 </LinkGrp>
</linkList>
```

Figure 7: Token alignment

when the bilingual BLIS corpus have been aligned at these levels. The main work for the encoding is to carry out the *id* pairing for each pair of source and target texts.

Figures 6 and 7 illustrate the alignment at both levels, respectively, in association with Figures 3 and 5. Within a <translation>, the *trans.loc* attribute specifies the location of the source of translation text and the *n* attribute indicates the designated order of the <align> elements in each <link>. The alignment is achieved by pairing up the unique *id* values in the annotation of primary data, each of which becomes the value for the *xlink:href* attribute in an <align>. If a sentence or token finds no counterpart in another language, i.e., it has no overt translation, the value “#xces:undefined” will be assigned to this attribute to specify such a case.

4.4 Web Browsing

BLIS is intended to serve as a bilingual searchable database for access from the Web. It is not really its concern whether its text display format is significantly different from the authorized print of the HK laws as in the loose-leaf edition.

The XML markup presented above provides a nice basis for rendering the authorized print of the laws in an exact way. Based on the markup, two

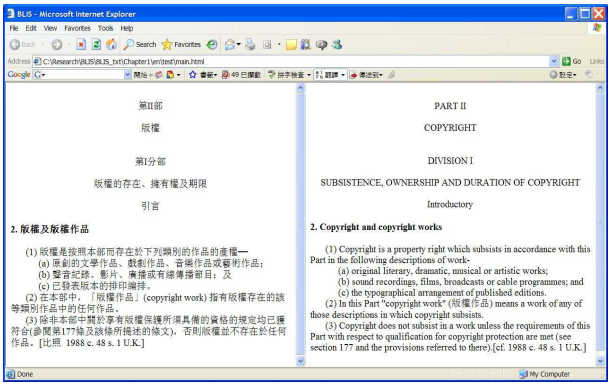


Figure 8: "Reprint" of bilingual HK laws on the Web

modes are provided for the Web browsing of the bilingual texts of the BLIS corpus. The first one displays the bilingual texts line by line in pairs. Notice that by a text line we refer to a string carrying no hard break as line terminator. In the BLIS corpus, it can be a subsubparagraph, subparagraph, a paragraph or even a subsection and may be printed in several lines by virtue of wrapping in the loose-leaf edition. This mode provides a convenient way for contrasting the parallel texts in the two languages in question. But a display mode for finer-grained bilingual browsing than this, e.g., for word alignment, is yet to be implemented.

The other browsing mode is intended to realize an accurate "reprint" of the loose-leaf version of the laws on the Web in two columns, each for a language. It serves both mono- and bi-lingual browsing. However, it is worth noting that such a Web display of the legal texts does not render the page breaking and line wrapping in the loose-leaf edition, because Web browsing allows scrolling up and down in a browsing window of various size. A sample display of BLIS text in this mode is illustrated in Figure 8, as a Web "reprint" for Figure 1.

5 Conclusion

The primary goal of corpus annotation is aimed at data capture at various linguistic levels. Without doubt, the existing annotation schemes serve this purpose very well, as exhibited in many researchers' recent practice in the field. However, it also remains a non-trivial task to apply and extend these schemes as necessary to represent the text structure of a document and the text hierarchy of a corpus for some other purposes, e.g., rendering the original print format of laws texts on the Web. As bilingual texts are concerned, this can also lay the necessary groundwork for bilingual legal drafting, which has been badly needed for long in a bilingual society like HK with two official languages.

In this paper, we have presented our recent work on XML encoding for the BLIS corpus of HK bilingual legislative texts following XCES. The linguistic annotation and text alignment are both encoded at the sentence and token levels using the available XCES schemas directly. To encode the BLIS hierarchy, however, we have to develop a set of intuitively readable tags on top of those from XCES for all discourse entities involved at various hierarchical levels, from the chapter down to the subsubparagraph. We adhere to the XCES guidelines and make the best of its schemas while making an effort to enhance the naturalness, and hence the readability, of the XML tag set to deal with the BLIS hierarchy. Based on this, our corpus encoding has allowed a nice rendering of the authorized print format of the loose-leaf edition of the bilingual HK laws for Web browsing.

For the future, we will extend our current work to encode the bilingual legal terminology in the BLIS corpus. We will also explore the possibility of developing utilities for bilingual legal drafting based on the XML markup that we have carried out.

References

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with Sara*. Edinburgh University Press.

Barthelemy, Francois, Pierre Bouiller, Pililppe Deschamp, Linda Kaouane, Abdelaziz Khajour, and Eric Villemont de la Clergerie. 2001. Tools and resources for tree adjoining grammars. In *ACL'01*, pages 63–70, Toulouse, July.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *ACL'91*, pages 169–176, Berkeley, California, June.

Burnard, Lou. 1999. Using SGML for linguistic analysis: the case of the BNC. *Markup Languages: Theory and Practice*, 1(2):31–51.

Erjavec, Tomaz, Jin-Dong Kim, Tomoko Otha, Yuka Tateisi, and Junichi Tsujii. 2003. Encoding biomedical resources in TEI: The case of the GENIA corpus. In *ACL'03*, pages 97–104, Sapporo, Japan, July.

Francis, W. Nelson and Henry Kucera. 1979. Brown Corpus Manual. Brown University, Providence.

Gale, William A. and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *ACL'91*, pages 177–184, Berkeley, California, June.

Gamper, Johann. 2000. A parallel corpus of Italian/German legal texts. In *LREC2000*, pages 531–538, Athens, Greece, June.

Hansen-Schirra, Slivia, Stella Neumann, and Mihaela Vela. 2006. Multi-dimensional annotation and alignment in an English-German translation corpus. In *NLPXML-2006*, pages 35–42, Trento, Italy, April.

Ide, Nancy. 1998. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In *LREC1998*, pages 463–70, Granada, Spain.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *LREC2000*, pages 825–830, Athens, Greece, May.

Ide, Nancy and Laurent Romary. 2001. A common framework for syntactic annotation. In *ACL'01*, pages 298–305, Toulouse, July.

Ide, Nancy and Keith Suderman. 2004. The American National Corpus first release. In *LREC2004*, pages 1681–1684, Lisbon, May.

Ide, Nancy and Keith Suderman. 2006. Integrating linguistics resources: The American National Corpus model. In *LREC2006*, Genoa, Italy, May.

Ide, Nancy and Jean Vronis. 1995. *The Text Encoding Initiative: Background and Context*. Kluwer Academic Publishers, Dordrecht.

ISO 8879. 1986. *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. ISO, Geneva.

Kit, Chunyu, Xiaoyue Liu, King Kui Sin, and Jonathan J. Webster. 2005. Harvesting the bitexts of the laws of Hong Kong from the Web. In *ALR-05*, pages 71–78, Jeju Island, October.

Kit, Chunyu, Jonathan J. Webster, King Kui Sin, Haihua Pan, and Heng Li. 2004. Clause alignment for bilingual hk legal texts: A lexical-based approach. *International Journal of Corpus Linguistics*, 9(1):29–51.

Klein, Dan and Christopher D. Manning. 2001. Parsing with Treebank grammars: Empirical bounds, theoretical models, and the structure of the Penn Treebank. In *ACL'01*, pages 338–345, Toulouse, France, July.

Lyding, Verena, Elena Chiocchetti, Gilles Serasset, and Francis Brunet-Manquat. 2006. The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 25–31, Sydney, Australia, July.

Ma, Wei-yun and Chu-Ren Huang. 2006. Uniform and effective tagging of a heterogeneous giga-word corpus. In *LREC2006*, Genoa, Italy, 24-28 May.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Muller, Christoph. 2006. Representing and accessing multi-level annotations in MMAX2. In *NLPXML-2006*, pages 89–92, Trento, Italy, April.

Palmer, Martha, Joseph Rosenzweig, and Scott Cotton. 2001. Automatic predicate argument analysis of the Penn Treebank. In *Proceedings of the first international conference on Human language technology research*, pages 1–5, San Francisco.

Ralf, Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC2006*, Genoa, Italy, 24-26 May.

Reppen, Randi and Nancy Ide. 2004. The American National Corpus: Overall goals and the first release. *Journal of English Linguistics*, 32(2):105–113.

Sperberg-McQueen, C. Michael and Lou Burnard, editors. 1994. *Guidelines for Electronic Text Encoding and Interchange* (*TEI P3*). Text Encoding Initiative, Oxford, Virginia, Brown.

Sperberg-McQueen, C. Michael and Lou Burnard, editors. 2007. *TEI Guidelines, P5*. Text Encoding Initiative, Oxford, Virginia, Brown. Available online at http://www.tei-c.org/P5/.

Suderman, Keith and Nancy Ide. 2006. Layering and merging linguistic annotations. In *NLPXML-2006*, pages 89–92, Trento, Italy, April.

Sun, Le, Song Xue, Weimin Qu, Xiaofeng Wang, and Yufang Sun. 2002. Constructing of a large-scale Chinese-English parallel coprus. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization*, pages 1–8, Taipei, Taiwan, August 31.

Wu, Dekai. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *ACL'94*, pages 80–87, Las Cruces, New Mexico, June.

Steps toward global interoperability for language resources

D. Terence Langendoen
dlangend@nsf.gov

Program Director for Linguistics and Cyberinfrastructure, National Science Foundation &
Professor Emeritus of Linguistics, University of Arizona, USA

For presentation at the First International Conference on Global Interoperability for Language Resources, City University of Hong Kong, 9-11 January 2008.

Abstract

This paper recounts my experience with the problem of interoperability of language resources over the past twenty years, during which time I have witnessed the progression in the development of markup languages from SGML to OWL and of tools for using them, which first raised the possibility of automatic semantic processing, and subsequently has brought the goal of achieving interoperability based on such processing closer.

I begin with my work with the Text Encoding Initiative (TEI). The TEI celebrated its 20th anniversary two months ago, and has many notable achievements in humanities computing to its credit, not the least of which is having pioneered in providing comparable (if not interoperable) resources for textual data and organization generally, and language data specifically. My role as chair of the TEI Analysis and Interpretation Working Group in the early 1990s was to help formulate document type definitions (DTDs) and prepare supporting documentation for a number of data structure varieties, including linguistically glossed text ("simple analytical markup"), typed feature structures, feature system declarations (effectively grammars expressed as bundles of feature structures), graphs and trees, and to assist other TEI working groups on problems ranging from alignment to the encoding of discontinuous spans to the logic of pointers.

My next major project was as co-principal investigator for the Electronic Metastructure for Endangered Language Data (E-MELD, NSF award BCS - 0094934) in the early 2000s, in which I led the initial development of the General Ontology for Linguistic Description (GOLD). The idea here is to provide a common, community-based resource that individual researchers or projects can use to identify the linguistic concepts used in their materials so as to support global sharing and interoperability of those materials. An example of the use of this resource is the Online Database of Interlinear Glossed Text (ODIN, BCS - 0411348) in which the concepts referred to in glossed text fragments in hundreds of different languages found on the Internet are identified by reference to GOLD.

Since May 2006, I have been working as a Program Director at the US National Science Foundation (NSF), in the Linguistics Program in the Division of Behavioral and Cognitive Science (BCS) in the Social, Behavioral & Economic Sciences Directorate (SBE) and in the Office of Cyberinfrastructure (OCI). As a member of the SBE Cyberinfrastructure Working Group and of the OCI Data Group, I have been helping manage several cyberinfrastructure solicitations including Community Based Data Interoperability Networks (INTEROP) and the new program in Documenting Endangered Languages (DEL), and am serving as a liaison to other US government agencies regarding cyberinfrastructure, including the National Endowment for the Humanities (NEH) and the National Institutes of Health (NIH).

My goal in this presentation is to suggest pathways for achieving global interoperability of language resources based on my experience in working in the field.

Multi-Argument Classification for Semantic Role Labeling

Chi-San Althon Lin
Department of Computer Science
Waikato University
Hamilton, New Zealand
cl123@cs.waikato.ac.nz

Tony C. Smith
Department of Computer Science
Waikato University
Hamilton, New Zealand
tcs@cs.waikato.ac.nz

Abstract

This paper describes a Multi-Argument Classification (MAC) approach to Semantic Role Labeling. The goal is to exploit dependencies between semantic roles by simultaneously classifying all arguments as a pattern. Argument identification, as a pre-processing stage, is carried at using the improved Predicate-Argument Recognition Algorithm (PARA) developed by Lin and Smith (2006). Results using standard evaluation metrics show that multi-argument classification, achieving 76.60 in F_1 measurement on WSJ 23, outperforms existing systems that use a single parse tree for the CoNLL 2005 shared task data. This paper also describes ways to significantly increase the speed of multi-argument classification, making it suitable for real-time language processing tasks that require semantic role labeling.

1 Introduction

The Conference on Natural Language Learning (CoNLL) has organized different shared tasks since 1999, including syntactic chunking, clause identification, name entity recognition, semantic role labeling (SRL), and multi-lingual dependency parsing. The goal of the problem of SRL as posed for CoNLL 2005 (Carreras and Marquez, 2005) is to recognize all the arguments of given predicates in a sentence and label them with appropriate semantic roles. Arguments related to a predicate are typically phrases in the sentence that form a relationship with the predicate. This relationship is called a semantic role. Generally speaking, SRL is a two step process (though some existing systems address it as a single task). Firstly, all arguments for a predicate must be identified with exact word spans—so-called *argument identification*. Secondly, these arguments must be labelled with correct semantic roles—referred to as *argument classification*.

Existing systems for semantic role labeling use machine learning methods to assign roles one-at-a-time to candidate arguments. There are several drawbacks to this general approach. First, more than one candidate can be assigned the same role, which is undesirable. Second, the search for each candidate argument is exponential with respect to the number of words in the sentence. Third, single-role assignment cannot take advantage of dependencies known to exist between semantic roles of predicate arguments, such as their relative juxtaposition. And fourth, execution times for existing algorithms are excessive, making them unsuitable for real-time use.

This paper seeks to obviate these problems by approaching semantic role labeling as a multi-argument classification process. It observes that the only valid arguments to a predicate are unembedded constituent phrases that do not overlap the predicate. Given that semantic role labeling occurs after parsing, this paper uses the Predicate-Argument Recognition Algorithm (PARA) by Lin and Smith (2006) that systematically traverses the parse tree when looking for arguments, thereby eliminating the vast majority of impossible candidates.

2 Syntax-Driven Argument Identification

Conventional argument identifiers, such as the one developed by Gildea and Palmer (2002), take all nodes in a parse tree, including each word in a sentence, as potential arguments (*pa*). Whether a potential argument is classified as a valid semantic argument depends on a probability estimation such as that given by Gildea and Palmer, (2002) or similar. Such a recognizer is a binary classifier, utilizing the distribution observed in the training data to learn how to predict future novel semantic arguments. Information from a parse tree is forwarded as features to the argument recognizer to help formulate a model to make correct predictions. The most-frequently used features for semantic arguments are the Path, Headword, Phrase type, and Predicate itself, as summarized in Table 1.

Predicate (*pr*) – The given predicate lemma (an uninflected, untensed verb).
Path (*path*) – The syntactic path through the parse tree from the constituent to the given predicate.
Head Word (*hw*) – The syntactic head of the phrase. (The head is normally simply the last noun of the rightmost subordinate noun phrase).

Table 1. Features used in semantic argument identification.

The statistical argument recognizer from Palmer et al. (2005) utilizes the following formula to estimate the probability of a potential argument:

$$P(pa \mid path, hw, predicate) = \lambda_1 * P(pa \mid path) + \lambda_2 * P(pa \mid path, predicate) + \lambda_3 * P(pa \mid hw, path)$$

where $\Sigma i\ \lambda_i = 1$.

Traditional argument recognizers have to spend time on each phrase and word to find possible semantic arguments. In order to reduce computational time, Xue and Palmer (2004) describe a pruning strategy to filter out constituents that are clearly not semantic arguments to the predicate. Then they classify the candidates derived from the pruning strategy as either semantic arguments or non-arguments. Finally they use a role classifier to label candidate arguments with semantic roles (Xue and Palmer, 2004). This pruning strategy has been widely used by systems in CoNLL2005 (Punyakanok et al., 2005; Tsai et al., 2005) to reduce training and testing time. Results (like Tsai et al. 2005) show this pruning strategy helps eliminate large portions of the training data (about 61% in Tsai et al. 2005) without sacrificing overall performance. Tsai et al. (2005) claim their systems with the pruning strategy achieve 93% of the correct arguments (or coverage) in training sets.

Generally speaking, valid arguments are non-overlapping and not embedded within each other. State-of-the-art syntactic parsers such as Collins (1999) or Charniak (2000) already solve the overlapping problem and their output provides an ideal structure for finding arguments. The residual problem is to select valid semantic arguments from these non-overlapping constituents of the parse trees. Cursory examination of hand-corrected parses reveals that upper-most nodes that do not include predicates are all valid potential arguments. PARA (Lin and Smith, 2006) was developed in accordance with this observation. The hypothesis is that upper-most nodes in the parse tree that do not include predicates are the potentially valid arguments and need not be rediscovered during argument identification.

PARA has been slightly modified so that it now ignores phrasal nodes that contain just punctuation symbols (an occasional error produced by automatic parsers). This turns out to improve PARA's performance quite significantly, as the following results for the CoNLL 2005 data demonstrate.

Approach	P	R	F_1
PARA-Imp	82.30	82.60	82.30
Moschitti	81.31	82.33	81.31
Palmer	81.30	80.62	80.96
Surdeanu	76.28	80.36	76.28
PARA	82.45	73.94	77.96

Table 2. Comparison of argument identification.

Table 2 shows comparison of argument identification on WSJ23 for four different approaches and the modified PARA (PARA-Imp). These results[1] are based on the official evaluation script[2] offered for the CoNLL shared tasks. The table shows the modification to PARA improves performance from

[1] All arguments are labeled with A0 except predicates.
[2] http://www.lsi.upc.edu/~srlconll/home.html

F_1: 77.96 to 82.60. The improved PARA also outperforms existing approaches using the same syntactic parses as input. It achieves the goal of a direct mapping from syntactic parses to unlabelled semantic arguments *without* the need for training by utilizing the output from a state-of-art parser, such as Charniak's (2000). The new PARA is fast and accurate, and can be used as a stand-alone preprocessor for the problem of Predicate-Argument Recognition or joined with other ML recognizers to increase the overall performance.

Utilizing the new PARA for argument identification, the following section introduces a new technique for multi-argument classification—one that outperforms existing systems in the SRL shared task, given single syntactic information (i.e. one parse tree per sentence).

3 Multi-Argument Classification

Approaches to argument classification are described in detail in the proceedings of CoNLL 2004 and CoNLL 2005 shared tasks. Many address argument classification using Machine Learning (ML) approaches; as with the SNoW learning architecture (Punyakanok et al., 2004, 2005), Support Vector Machines (Moschitti et al., 2005), and so on. The general trend is to try to increase performance by adding more features.

In contrast, this paper applies the concept of Multi-Argument Classification (MAC) to achieve better performance without additional features. MAC is based on the idea of exploiting relationships between roles in predicate-argument structures (e.g. [A0 V A1], [A1 V], etc). Such a relationship is called a ***semantic role dependency***. The relationship in the predicate-argument structure exhibits **semantic role dependency** manifest in the sequential order, count and juxtaposition of different core roles (like A0, or A1) in the predicate-argument list. Generally speaking, there is only one core role in each predicate structure[3]. This can serve as useful information for role classification, as demonstrated in the following classification model.

[3] There is rare situation happened with more than one core role.

3.1 Classification Model

Gildea and Jurafsky (2002) calculate the probability of the optimal role assignment r^* for each sentence as follows.

$$r^* = \text{argmax}_{r_{1\_n}} P(\{r_{1..n}\} \mid predicate) \prod_i \frac{P(r_i \mid \{f_i\}, predicate)}{P(r_i \mid predicate)}$$

$P(r_i \mid \{f_i\}, predicate)$ is the probability of a constituent's role given the above features for the constituent and the predicate. More detail is given in Gildea and Jurasky (2002).

This is a typical ML approach for maximizing the probability of the optimal role assignment to assign roles for each sentence without utilizing role dependency learned from the training data. In Multi-Argument Classification, the optimal probability applied with the role dependency relationship learned from the training data is as follows.

$$r^* = \underset{\{r_{1...n}\}}{\text{argmax}} P(\{r_{1..n}\} \mid predicate) \prod_i \frac{P(\{r_i\} \mid \{f_i\}, predicate)}{P(\{r_i\} \mid predicate)}$$

where $\{r_{1...n}\}$ is a sequential role list learned from the training data, $\{r_i\}$ is the j-th role in $\{r_{1...n}\}$, $P(\{r_{1...n}\} \mid predicate)$ represents the probability of an overall assignment of the role list $\{r_{1...n}\}$ to each of the n constituents or semantic arguments of a sentence, given the *predicate* and the various features $\{f_i\}$of each of the constituents.

The role list $\{r_{1...n}\}$ denotes there are n arguments in a test sentence; but the number of arguments in any training sentence may vary. To compare instances of different lengths, we add a mapping function to convert the role list $\{r_{1...m}\}$ of a training sentence to the role list $\{r_{1...n}\}$ of a test sentence as follows:

$$M: \{r_{1...m}\}_j \rightarrow \{r_{1...n}\}$$

where $\{r_{1...m}\}_j$ is the role list with m arguments of a training sentence j and $\{r_{1...n}\}$ is the role list with n arguments of the test (i.e. query) sentence. The basic principle of this mapping function is to map m arguments of a training sentence to n arguments of the query.

By replacing $\{r_{1...n}\}$ with $M\{r_{1...m}\}_j$ and $\{r_i\}$ with $\{r_{ki}\}$ in the previous formula, the probability formula for MAC is shown as follows.

$$r^* = \underset{M\{r_{1...m}\}_j}{\text{argmax}} P(\{r_{1..n}\} \mid predicate) \prod_i \frac{P(\{r_{kj}\} \mid \{f_i\}, predicate)}{P(\{r_{kj}\} \mid predicate)}$$

where $M\{r_{1...m}\}_j$ is the role list generated by the mapping function M from the *j*-th training sentences with *m* arguments of the training data to the role list $\{r_{1...n}\}$ for the test sentence with *n* arguments, and $\{r_{ki}\}$ denotes the *k*-th role of $\{r_{1...m}\}_j$ ($1 <= k <= m$) corresponding to the *i*-th argument of $\{r_{1...n}\}$. The details of the mapping algorithm are described in the next section.

3.2 Mapping Algorithm

There are four considerations essential to the function that maps a knowledge pattern learned from the training data to a new query sentence: *i)* where to start matching two patterns; *ii)* how to deal with different numbers of arguments between the knowledge and query patterns, *iii)* how to compute similarity between an argument in a knowledge pattern and an argument in a query pattern, and *iv)* how to measure the quality of the matching.

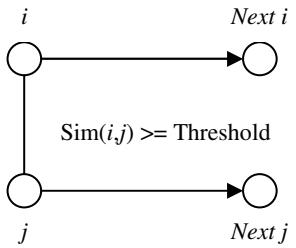
i) Where to start

The first consideration is solved by looking for the most common instance in a knowledge pattern and a query one, given the predicate. In this discussion, an instance is an argument in a predicate-argument structure.

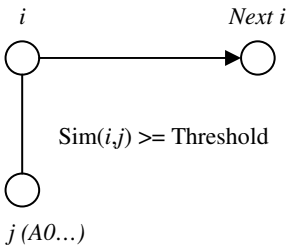
ii) Mapping of different arguments

Empirically there is very low coverage or recall (about 0.46) to match query sentences with training sentences that have the same number of arguments. This paper proposes an alternative way to increase the coverage. The principle is based on semantic role dependency, in which core roles (like A0 or A1) are regarded as *more essential* than adjuncts (like AM-TMP, or AM-LOC). We need to estimate similarity between an argument (i.e. instance) in a knowledge pattern and an argument in the query. If two instances (one in the knowledge pattern and the other in the query) are considered highly similar (Case 1), we can try to match the next instances in both patterns. If two instances are not similar, there are two kinds of situation. One is to match the current instance in the knowledge pattern with the next instance in the query (Case 2). The other is to match the next instance in the knowledge pattern with the current instance in the query (Case 3). The two final circumstances are unmatched instances in the query (Case 4), and unmatched instances in the knowledge pattern (Case 5). These five cases are more formally described as follows:

Case 1: if there exists a query instance *i* and a corresponding knowledge instance *j*, and both instances are similar (or their similarity is no less than a threshold), try to match the next instance in the knowledge pattern with the next instance in the query. This is the case when two instances are considered highly similar, then try to match the next instances in the query and knowledge patterns.

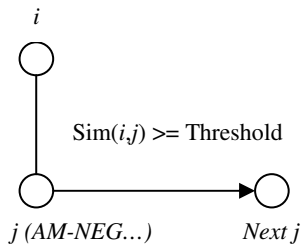


Case 2: if there exists a query instance *i* and a corresponding knowledge instance *j*, both instances are not similar (or their similarity is below a threshold) and the role of the knowledge instance j appears to be one of the core roles (i.e. A0 to A5 and AA), as opposed to a non-core or adjunctive role, try to match the current instance in this knowledge pattern with the next instance in the query. The reason to keep the current knowledge instance is to try to increase the coverage. It is rare to have two patterns match exactly due to inherent data sparseness. For example, a query sentence "They will come" and a training sentence, "They come" is not matched due to different number of arguments. But they can be considered highly similar if the second argument "will" in the query sentence is skipped during matching. Such a skipped argument can be labeled latter by other approaches.



Case 3: if there exists a query instance *i* and a corresponding knowledge instance *j*, both instances are not similar and the role of the instance *j* in the knowledge pattern appears to be a non-core label (e.g. AM-MOD AM-NEG or AM-DIS), try to match the next instance in the knowledge pattern with the current instance in the query. Such non-

core roles are optional to a query pattern—which is to say that not all sentences have them. This means they can be skipped. This is the complement situation to Case 2 where all non-core or adjunctive roles in the knowledge pattern can be skipped in order to increase coverage.



Case 4: if there does not exist an argument *j* in the knowledge pattern, keep a default probability to query instance *i* to avoid zero frequency.

Case 5: skip all extra corresponding knowledge arguments.

After mapping, all patterns from the pattern base are compared to role lists with the same number of arguments as the query. It remains now to measure the similarity of each argument in the query role list with each corresponding argument in the role list of each pattern from the knowledge base.

iii) Similarity Function

The calculation of similarity between an instance in the knowledge pattern and its corresponding instance in the query is based on the feature space. The distance between two points (i.e. instances in the feature space) is estimated by Euclidean distance as follows.

Distance metric (Euclidean distance):

$$D(x_i, x_j) = \sqrt{\Sigma(a_r(x_i))-a_r(x_j))^2}$$

for r =1 to *n* (i.e. *n* different classifications), where $a_r(x)$ means the r-th feature of an instance x. (Features used are described in Section 3.5.) If instances, x_i and x_j, are identical, then $D(x_i , x_j)=0$; Otherwise, $D(x_i , x_j)$ represents the vector distance between x_i and x_j. x_i is an instance in the query and x_j is an instance in the knowledge pattern.

Therefore the similarity function is defined as Sim(i, j) = (number of features - D(i, j)) / (number of features)

If all features are the same between two instances, D(i, j) is zero and Sim(i, j) is 1.0. If there are different features between two instances (for example two features are not the same), the score of similarity by Sim(i, j) will be less than 1.0. For example, if there are five features for calculation and two different, D(i, j) is two and Sim(i, j) is 0.6, which is (5 – 2) / 5. The threshold utilized in the first three cases is initially set to 1.0, which means all features in the knowledge instance must be the same with the ones of the query pattern. The threshold value was arrived at through trial and error.

iv) What is the quality estimation

The fourth issue mentioned early in this section is to find the quality of a match between a pattern in a training sentence and a query pattern in the query sentence by the formula given in Section 3.1, except that argument maximization is not used.

Once all quality probabilities for patterns in the training data are calculated, the system selects the pattern from the training sentence with the highest quality probability.

3.3 Unlabeled Arguments

MA is designed for matching two patterns with different arguments. It helps to increase the overall coverage from 0.46 (if only marching patterns with the same number of arguments) to 0.78. This is still not good enough compared to statistical singular-argument classifiers. The cause of low coverage is sparseness of data. For example, a skipped argument like "will" in the query of Case 2 can be labeled by other approaches (i.e. existing classifiers). Thus we propose a simple argument labeler to fill unlabeled arguments.

$$\underset{r}{\text{argmax}}\ P(r \mid \{f\}, \textit{predicate})$$

where $P(r \mid \{f\}, \textit{predicate})$ represents the probability of an assignment of role *r* (excluding any core role that already appears in the label list to avoid duplication of core roles) to each of the unlabeled arguments of a sentence after MA, given the *predicate* and the features {*f*} (including headword, distance, voice, preposition, phrase type and path) of the argument. By handling unmatched arguments with this simple argument labeler, the recall rises from 0.78 to 0.86.

3.4 Complete PM Model

The complete model for Pattern-Matching (PM) is thus a combination of MAC and SAC. PM tries to find all suitable patterns from the training data using the mapping algorithm described in Section 3.2,

selects the best one from the pattern base according to the quality probabilities from the mapping algorithm using MAC, and classifies any unlabelled arguments in the best pattern with SAC like a simple argument labeler in Section 3.3.

Procedure of Pattern-Matching with SAC

For all knowledge patterns
    apply Mapping Algorithm for the query and knowledge patterns

Select the best knowledge pattern according to their quality probabilities

Use SAC to classify the unlabelled arguments

Figure 1. Procedure of the PM model and SAC.

The goal of selection is to find the knowledge patterns with the highest Quality, calculated by MA described in Section 3.2. The procedure for PM is shown in Figure 1.

In the testing stage for the system, PARA is used as an argument recognizer to identify predicates and arguments related to predicates. It forwards the predicates and their arguments to classification. Argument classification in the system includes two role classifiers, a multi-argument classifier, Pattern-Matching (PM) described and a statistical singular argument classifier modified from Palmer et al. (2005). The modification includes two extra features, preposition and distance described as follows.

3.5 Features

Features used in this paper are predicate, voice, phrase type, distance, headword, path and preposition, as shown in Figure 2.

Predicate – The given predicate lemma.
Voice – Whether the predicate is realized as an active or passive construction.
Phrase Type – The syntactic category (NP, PP, S, etc.) of the phrase corresponding to the semantic argument.
Distance – The relative displacement from the predicate, measured in intervening constituents (negative if the constituent is to the left of or prior to, positive if it is to the right of or after, the predicate).
Head Word – The syntactic head of the phrase.
Path – The syntactic path through the parse tree, from the parse constituent to the predicate being classified.
Preposition – The preposition of an argument in a PP such as *during, at, with*, and so on.

Figure 2. Features used for experimentation..

4 Experiments and Results

Data used in this chapter is that released on March 2005 for CoNLL-2005 [4], which includes Wall Street Journal sections with Charniak's (2000) and Collins' (1999) parse-trees. Charniak's parse tree is accepted as input to the system due to its better performance on WSJ (Carreras and Marquez, 2005). Evaluation is carried out using the official evaluation script from CoNLL 2005, *srl-eval.pl* which provides precision, recall and F_1 measure of the predicated arguments. Predicates are given in the CoNLL shared tasks.

Table 3 shows the results for several approaches, when used with known arguments (i.e. the systems are given the correct arguments for role classification). All training data (WSJ02-21) with Charniak's parses are included. The modified version of the classifier from Palmer et al. (2005) (*Palmer-Imp)* provides 85.59 in F_1 and the performance of the basic model (*PM without Palmer*) estimation is F_1: 1.16 improved compared to *Palmer* itself. The complete model (*PM*), combined with *Palmer*, achieves the best results on Precision (88.89), Recall (87.65), and F_1 measurement (88.27) and offers the best solution on all test datasets compared to *Palmer-Imp*. It suggests *PM*, utilizing role dependencies existing in semantic roles, helps to increase F_1 by 3.0 over *Palmer-Imp*.

Table 4 shows the result using all features (*ALL*), and the contribution of each feature in Precision (P), Recall (R), and F_1 measurements.

Approach	P	R	F1
Pamler-Imp	85.53	85.65	85.59
PM without Palmer-Imp	87.67	85.85	86.75
PM	**88.89**	**87.65**	**88.27**

Table 3. Results obtained by different algorithms on WSJ Section 24 with known arguments.

[4] http://www.lsi.upc.edu/~srlconll/soft.html

	P	R	F1
ALL	88.89	87.65	88.27
- Voice	88.52	87.02	87.77
- Head Word	85.52	83.93	84.72
- Phrase Type	85.21	**83.03**	84.11
- Preposition	**84.77**	**83.03**	**83.89**
- Distance	88.81	87.56	88.18
- Path	87.12	85.89	86.50

Table 4. Contribution of each feature on WSJ 24, with known arguments.

Test dataset	P	R	F1
WSJ 24	75.88	72.98	74.40
WSJ 23	78.04	75.20	76.60
Brown	69.33	63.44	66.25

Table 5. Results for different training datasets on WSJ 24 with Charniak's parses and PARA-Imp.

Test WSJ23	**Precision**	**Recall**	**$F_{\beta=1}$**
Overall	78.04%	75.20%	76.60
A0	84.31%	85.18%	84.74
A1	78.86%	76.98%	77.91
A2	70.83%	61.26%	65.70
A3	68.84%	54.91%	61.09
A4	66.67%	62.75%	64.65
A5	100.00%	60.00%	75.00
AM-ADV	59.07%	55.34%	57.14
AM-CAU	64.91%	50.68%	56.92
AM-DIR	35.53%	31.76%	33.54
AM-DIS	76.25%	76.25%	76.25
AM-EXT	50.00%	37.50%	42.86
AM-LOC	62.54%	51.52%	56.50
AM-MNR	59.33%	51.74%	55.28
AM-MOD	97.42%	95.83%	96.61
AM-NEG	95.18%	94.35%	94.76
AM-PNC	46.39%	39.13%	42.45
AM-PRD	0.00%	0.00%	0.00
AM-REC	0.00%	0.00%	0.00
AM-TMP	73.58%	72.49%	73.03
R-A0	85.84%	86.61%	86.22
R-A1	80.28%	73.08%	76.51
R-A2	80.00%	50.00%	61.54
R-A3	0.00%	0.00%	0.00
R-A4	0.00%	0.00%	0.00
R-AM-ADV	0.00%	0.00%	0.00
R-AM-CAU	0.00%	0.00%	0.00
R-AM-EXT	0.00%	0.00%	0.00
R-AM-LOC	73.68%	66.67%	70.00
R-AM-MNR	25.00%	16.67%	20.00
R-AM-TMP	62.69%	80.77%	70.59

Table 6. Details for each semantic role on WSJ 23, with Charniak's parses and PARA.

Phrase Type, Preposition, and Head word are the three features whose removal decreases the performance of the complete system by a large amount. The distance feature plays a key role in overall performance of *Palmer-Imp* but is the least influential in *PM* because of the usage of multi-argument classification. When using *PM*, the related distance is implicitly included when matching two patterns. The path feature is the fourth most influential factor on performance for role classification, and the voice feature has the least detrimental effect, along with the distance feature, on the performance of this system. Both features (path and voice) have the same influence in *PM* and *Palmer-Imp*.

Table 5 shows performance (on WSJ 24, WSJ 23 and the Brown corpus) of the complete model (PM) using auto parses (Charniak's parser) and PARA as the pre-processor to recognize all related arguments. It also shows the results on WSJ 23 are about F_1:2.0 better than that by WSJ 24. This increase is because the performance by PARA on WSJ 23 is about F_1:2.0 better than WSJ 24. The results on the Brown corpus show the performance drops by more than 10 points in F_1 compared to WSJ 23. This is caused by propagating process-errors described in Carreras and Marquez, (2005). Table 5 also shows such errors affect results even more in the domain of the Brown corpus. Another area for future work is to look for ways to minimize the impact of different domains.

The results on WSJ 23 for each role are shown in Table 6. Generally speaking, performance on core roles is better than on adjuncts, except for the modal, and negation tags. This is because there are more training examples for core roles than for adjuncts.

Experimental results show that execution times for *PM* and *Palmer-Imp* are about 3.0 and 0.8 seconds per sentence respectively. To increase speed, we introduce a controlling strategy called the Maximum Suitable Pattern (MSP) number. MSP limits how many suitable patterns must be found for a query pattern before searching/comparing can stop. The MSP formula is:

$$r^* = \underset{\substack{\text{Suitable(j)} <= \text{MSP} \\ M\{r_{1\ldots m}\}_j}}{\operatorname{argmax}} P(\{r_{1..n}\} \mid predicate) \prod_i \frac{P(\{r_{kj}\} \mid \{f_i\}, predicate)}{P(\{r_{kj}\} \mid predicate)}$$

where Suitable(j) denotes the number of suitable knowledge patterns found.

Once PM has found enough suitable patterns (Suitable(j) > MSP), PM stops matching knowledge patterns in the pattern base. A knowledge pattern with at least one instance that has similarity probability greater than the threshold is defined as a suitable one.

Table 7 shows different results for various values of Maximum-Suitable Pattern (MSP) and suggests no improvement after 100 matches. Note that all accuracy differences appear insignificant, but the execution time per sentence (T) increases as the MSP value does, suggesting an MSP between 10 and 20. All execution time are calculated based on a P4 3.0 GHz CPU and 1G RAM Linux machine.

MSP	P	R	F1	T
30000	89.68	89.20	89.44	2.949
10000	89.68	89.20	89.44	2.943
1000	89.67	89.17	89.42	2.433
100	89.71	89.39	89.55	1.235
50	89.73	89.39	89.56	1.035
20	89.84	89.54	89.69	0.858
10	89.78	89.34	89.56	0.788
2	89.13	88.58	88.86	0.609
1	89.00	88.32	88.66	0.591

Table 7. Results for different MSP values obtained on WSJ 23, with known arguments

System	*P*	*R*	F_1	*NoF*
PARA+PM	78.04	75.20	76.60	**7**
Surdeanu	80.32	72.95	76.46	31
Tsai	82.77	70.90	76.38	25
Moschitti	76.55	75.24	75.89	14
PARA+Palmer-Imp	71.18	70.90	73.49	7

Table 8. Results for different systems on WSJ 23 listed in the CoNLL 2005 shared task.

Table 8 shows comparative results for various systems using the same input. Surdeanu et al. (2005), Tsai et al. (2005), and Moschitti et al. (2005) are systems only using Charniak's parses listed in CoNLL 2005 shared task. The modified system (*PARA+Palmer_Imp*) is the combination of PARA and *Palmer-Modified.* Even using fewer features, the combination of *PARA* and *PM* offers a more accurate system for SRL compared to systems using the same input. It also becomes one of the top-performing systems in the CoNLL 2005 shared task compared to systems using far more features and multiple parses. It suggests that exploiting role dependencies helps improve accuracy in SRL.

References

Carreras, X. and Marquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Charniak, E. (2000). A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL-2000*.

Collins, M. (1999). Head-Driven Statistical Models for Natural Language Parsing. *PhD Dissertation, University of Pennsylvania*.

Gildea, D. and Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics, 28(3):245-288*.

Gildea, D. and Palmer, M. (2002). The Necessity of Parsing for Predicate Argument Recognition . In *Proceedings of ACL 2002, Philadelphia, USA*.

Lin, C.S. A. and Smith, T. C. (2006). A Tree-based Algorithm for Predicate-Argument Recognition. In *Bulletin of Association for Computing Machinery New Zealand (ACM_NZ), volumn 2, issue 1*.

Moschitti, A., Giuglea, A.-M., Coppola, B., and Basili, R. (2005). Semantic role labeling using support vector machines. In *Proceedings of CoNLL-2005*.

Palmer, M., Gildea, D., abd Kingsbury, P., (2005). The Propostin Bank: An Annotated Corpus of Semantic Roles. In *Proceedings of ACL: Volume 31, Number 1. p72-105*.

Punyakanok, V., Koomen, P., Roth, D., and Yih, W. T. (2005). Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*.

Surdeanu, M. and Turmo, J. (2005). Semantic role labeling using complete syntactic analysis. In *Proceedings of CoNLL-2005*.

Tsai, T.-H., Wu, C.-W., Lin, Y.-C. and Hsu, W.-L. (2005). Exploiting Full Parsing Information to Label Semantic Roles Using an Ensemble of ME and SVM via Integer Linear Programming. In *Proceedings of CoNLL 2005*.

Xue, N. and Palmer, M. (2004). Calibrating features for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

The New FrameNet Desktop: A Usage Scenario for Slovenian

Birte Lönneker-Rodman, Collin Baker, Jisup Hong
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA
{loenneke,collinb,jhong}@icsi.berkeley.edu

Abstract

The FrameNet Desktop application supports frame-semantic lexicon creation and corpus annotation. It is primarily intended for use with the English language at the FrameNet project, but a recent revised version facilitates easier distribution, installation and adaptation to new projects. This paper reports on a case study where the new FrameNet Desktop is used with Slovenian data, in a scenario involving lexicon import from and annotation export to an electronic dictionary.

1 Introduction

FrameNet is a rich lexical semantic resource for English supported by manual annotation of electronic corpora (Ruppenhofer et al., 2006). The FrameNet project has developed the FrameNet Desktop, an application initially intended for frame-semantic lexicon creation and annotation of English texts at the project site. But with a growing number of projects for other languages referring to the FrameNet model, the FrameNet Desktop has recently undergone various changes in architecture and functionality. The aim of these efforts is to make the tool easier to distribute and reuse for purposes including manual annotation in other languages, as well as viewing, summarizing and correcting the output of automatic frame-semantic annotation.

This paper presents a case study aimed at adapting FrameNet Desktop to a new language (Slovenian), and at the same time exchanging data between the FrameNet tool and an electronic dictionary. While the data extracted from the dictionary for import into the Desktop application consists in purely lexical and morphologic data (words, multiwords, and word forms), FrameNet Desktop output contains rich semantic information by associating lexical units with semantic frames, and annotated sentences illustrating their usage in context. These sentences can be fed back into the dictionary, where they will serve as usage examples.

The remainder of the paper is subdivided into two main sections. First, we will provide an overview of the FrameNet Desktop application and recent developments (Section 2). Second, a case study on utilizing the revised FrameNet Desktop for Slovenian and interchanging data between it and an electronic dictionary will be presented (Section 3). A brief summary and outlook in Section 4 concludes the paper.

2 FrameNet Desktop

After explaining FrameNet Desktop core functionality by way of an example (2.1), projects for other languages making use of earlier versions of the Desktop software (2.2) as well as its general architecture (2.3) will be described. Finally, recent developments of the tool are presented in Subsection 2.4.

2.1 Core functionality

The FrameNet Desktop is a suite of tools for (a) entering and editing frame descriptions and lists of lexical units, (b) extracting and annotating representative sentences from a corpus for *lexicographic annotation*, or annotating every target word in a given corpus during *full-text annotation*, and (c) organiz-

ing and displaying the results. Annotating target words in the FrameNet Desktop means identifying them as targets, selecting the lexical unit and frame they instantiate, and attaching frame-relevant labels to other constituents of the same sentence.

On the example of the ATTACHING frame, (Fillmore et al., 2003) gives a detailed overview of the annotation process within FrameNet Desktop, which illustrates many functions of the tool. In what follows, we will narrow down their examples to the discussion of a few core functions.

Frame and lexical unit creation. Creating a semantic frame involves describing (in English prose) the type of situation or happening it refers to, and drawing a list of words with senses that can be explained with reference to this frame. For instance, in the situation of the ATTACHING frame, somebody attaches (or affixes or joins) one thing to another thing, using some kind of connector. This is reflected in the frame description. A frame also includes a list of participant roles or aspects of the situation, called frame elements (FEs); for ATTACHING, the most prominent ones ("Core FEs") are: **Agent** (the person who brings about the attaching), **Item** (an object affixed to a larger, more stable Goal), the **Goal**, and a **Connector** (the bond). Word senses explained in terms of a semantic frame are *lexical units* (LUs) of that frame. The list of English verb LUs belonging to ATTACHING includes *append*, *attach*, *connect*, and *tie*, among others. The FrameNet Desktop thus provides a means to create frames along with their informal description and frame elements, as well as lexical units, which are defined as the combination of a frame and a lemma (a lemma in FrameNet is a complex lexical entity combining information on lexeme, part of speech, and word forms; see e.g. (Baker et al., 2003)).

Annotation of LU occurrences. For lexicographic and Natural Language Processing purposes, it is of interest to know how frame elements are linguistically expressed in sentences containing frame LUs. Therefore, FrameNet Desktop facilitates the annotation of sentences relative to one lexical unit, the *target*. Whole constituents, dependents of the target word within the selected sentence, are annotated on separate layers for frame element, grammatical function, and phrase type, among others. To the present discussion, the FE layer is most central. It has been illustrated by (Fillmore et al., 2003) with the annotation of an occurrence of the lexical unit *tie.v* in sentence (1).

(1) Apparently the healer would tie a black thread round the horse's ankle, and it usually worked.

On the FE annotation layer for the target *tie* in this sentence, the FE **Agent** is assigned to *the healer*, **Item** to *a black thread*, and **Goal** to *round the horse's ankle*. FrameNet Desktop provides a graphical interface for this kind of work, where words and sequences of words are selected, and frame elements from a frame-specific list assigned to them, with a few mouse clicks.

2.2 FrameNet Desktop for other languages

Initially, the FrameNet Desktop was designed for work on English. Since then, it has been adapted by various sites setting up FrameNets for different languages. To the knowledge of the authors, variants of earlier FrameNet Desktop versions are in use at Spanish FrameNet (Subirats and Sato, 2004; Subirats, Under review), Japanese FrameNet (Ohara et al., 2004), and German FrameNet[1]. More or less important changes to the FrameNet Desktop software were carried out at the respective project sites. Extent and nature of the changes depend on the resources and aims of the different projects, as well as on language-specific properties, such as differences in character encoding or morpho-syntactic structure.

2.3 FrameNet Desktop system architecture

The FrameNet Desktop is a J2EE-based client/server application system using MySQL for data storage. The server side requires a networked server with a Unix-like operating system, such as Linux or Mac OS X. To start up a new FrameNet Desktop project, a MySQL database must be created and seeded with the necessary table structures (Baker et al., 2003) as well as with all data the project chooses to take over from the original FrameNet database (this might include frame and frame element definitions or information about available annotation labels). This MySQL database serves as the data

[1] `http://gframenet.gmc.utexas.edu/` [September 28, 2007]

source for a JBoss application server, which mediates all interaction with FrameNet Desktop clients through FrameNet-developed Enterprise JavaBeans (EJBs). The JBoss server also ensures transactional integrity in a multiuser environment. Any number of Desktop clients, within which the manual lexicographic work and annotation finally takes place, can be run from any machine with network access to the server.

2.4 Recent developments

At the FrameNet project, a new version of the FrameNet Desktop software has been deployed that compiles and runs on Java 5 (aka. Java 1.5) and the JBoss 4.0.5.GA J2EE application server. Besides improvement in performance thanks to the newer versions of Java and JBoss, extensive refactoring of the code-base promises to facilitate ongoing internal development as well as make source code distribution to outside groups feasible. Changes range from linearizing package-to-package dependencies and establishing clearer boundaries between the server and client tiers, to rearranging the source directory structure to accommodate popular IDEs like Eclipse and NetBeans, to integrating JBoss and MySQL configuration into the build infrastructure. In the new system, for example, Ant commands can be used to create a new database and seed it with the necessary table structure and to start and stop MySQL and JBoss. These changes will allow FrameNet to distribute Desktop source code along with a fully-functional development and testing configuration that can be easily adapted by an outside project for different languages.

Currently, a revised XML format is being designed and implemented that can be used for both input and output for the FrameNet Desktop software. This is expected to improve data interchange with other applications. Usage scenarios include importing the output of automatic frame-semantic annotation into the Desktop application, manually correcting it therein, and exporting the corrected version for retraining the statistical data model used by the automatic labeler. Another example of how annotated texts exported from FrameNet Desktop can be used within other applications is provided in the case study below (Section 3).

Finally, functionality and user-friendliness of the new FrameNet Desktop have been improved. This includes full Unicode support as well as a character-by-character selection mode. The latter can now be toggled for each session and allows the annotation of units below the word level, which can be useful when working with languages rich in inflection and/or compounding.

3 Case study: A new usage scenario for FrameNet Desktop

This section describes how the FrameNet Desktop is set up and used to work with a new language, Slovenian. To the authors' knowledge, no FrameNet project has been undertaken for this small Slavic language before. A usage scenario will be presented, involving data interchange between the Desktop application and an electronic dictionary covering Slovenian. The intended goal is to collect frame-semantically annotated sentences illustrating the usage of lexemes covered by the dictionary. The sentences can then be displayed along with the main dictionary entries, to provide the user with examples of the word used in context. An overview of the entire process is given in Figure 1.

The dictionary – provider of lexical information and recipient of annotated example sentences – is briefly described in 3.1, before the steps relevant to FrameNet Desktop will be covered: importing language-independent information (3.2), lexicon data (3.3), and corpus files (3.4), annotating text (3.5), and exporting the annotated corpus for use in the target application (3.6).

3.1 The electronic dictionary

The electronic dictionary used in this case study is *Online SLO-DE-SLO*, a Slovenian-German/German-Slovenian online dictionary[2] (Lönneker and Jakopin, 2003; Jakopin and Lönneker, 2004). The motivation behind *Online SLO-DE-SLO* is to create a lexical resource useful to both human users and Natural Language Processing. The dictionary contains over 8,800 word and multi-word correspondences, as well as 2,150 bilingual usage examples. It records, on average, more than 80,000 requests per month. Data is stored

[2] `http://webapp.rrz.uni-hamburg.de/~slowenisch/` [September 28, 2007].

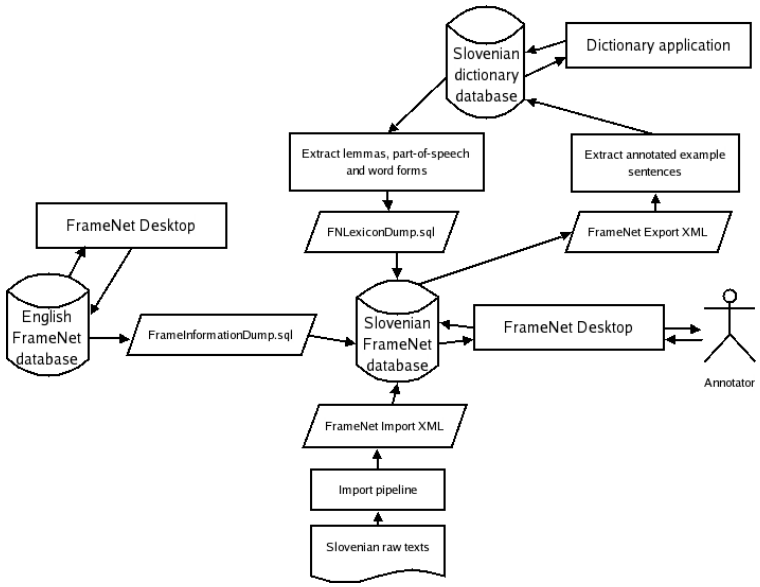


Figure 1: Data interchange between electronic dictionary and FrameNet Desktop.

in a MySQL database, and HTML forms and Perl scripts are used to display dictionary entries in the Web interface.

Whereas the core vocabulary entered into the dictionary was manually drawn from textbooks of both languages, the dictionary has recently been extended by entries derived from electronically available Slovenian texts covering various domains (short stories, cooking, transportation, and transcripts or written drafts of public speeches). These text corpora will also be used for frame-semantic annotation with FrameNet Desktop.

3.2 Importing language-independent information into FrameNet Desktop

Besides the empty table structure of the FrameNet database, new FrameNet-related projects might also take over from the English FrameNet database what they consider being reusable, language- or project-independent, information. Following the example of previous non-English FrameNets (see 2.2 above), we decided to import all frame and frame element definitions from the English FrameNet database by producing an SQL dump file (`FrameInformationDump.sql` in Figure 1) and loading it into the Slovenian FrameNet database.

This makes available the full list of English frames in FrameNet Desktop, where Slovenian lexical units can be attached to them, and their frame elements used during annotation. Frames can also be modified within FrameNet Desktop; thereby, language-specific differences in frames or frame elements can be accommodated.

3.3 Importing lexicon data

When adding a lexical unit to a frame, the FrameNet Desktop allows the user to create a new lemma, which will be added to the lexicon part of the FrameNet database, together with its corresponding lexeme(s) and word forms, also to be entered manually. Alternatively, the Desktop application allows the user to choose an existing lemma from a list, and to add it to a frame as a new LU.

When starting a FrameNet project for a new language, it might not be desired to start with a completely empty lexicon, and being forced to manually enter all lexemes and word forms; this is especially true for morphologically rich languages, such as Slovenian. A more practical approach is to populate the lexicon part of the FrameNet database with lemmas, lexemes and word forms from an external source, before starting LU creation in the FrameNet Desktop, so that existing lemmas can always be selected from a list.

In our experiment, the relevant lexicon data for Slovenian has been extracted from the Slovenian-

German dictionary. A series of SQL commands is used to prepare tables in the right format and conflate redundant information. For example, the dictionary contains ordered lists of all word forms of a lexeme, which are further described by morphological information. FrameNet Desktop does not store information on individual word forms; therefore, a simple set of distinct forms pertaining to a lexeme is extracted for import into the annotation tool. The file produced by the SQL commands is an SQL dump file (`FNLexiconDump.sql` in Figure 1), which can easily be imported into an empty FrameNet database. A further advantage of this approach is that dictionary entries and lexemes in the FrameNet database are implicitly linked via identical IDs, thereby increasing interoperability.

Online SLO-DE-SLO dictionary entries exhaustively cover the vocabulary of all corpus texts currently considered for import into the Slovenian FrameNet Desktop (see 3.1 above). Therefore, it is not necessary to add FrameNet lexicon entries in the case study. In the future, it would be desirable to update the Slovenian FrameNet lexicon tables by importing additional data from the dictionary, as the latter offers automatic word form generation. Taking over new lexical data from the dictionary instead of creating it in FrameNet Desktop would thus increase data consistency and minimize the risk of misspelled or overlooked word forms.

3.4 Importing corpus files

FrameNet XML to be loaded into FrameNet Desktop for full-text annotation is produced by the import pipeline (`ImportFullText`, a Desktop utility implemented in Java) from a text file encoded in UTF-8 character set and provided with paragraph boundary tags. For English, text import comprises the following steps:

1. sentence boundary detection
2. Named Entity Recognition (NER)
3. part-of-speech tagging and lemmatization
4. recognition of tokens that should not receive frame semantic annotation (e.g., punctuation, proper names)
5. XML formatting.

Import of Slovenian texts starts at step 3 of the pipeline, thus presupposing a text file provided with both paragraph and sentence boundary tags. The NER step is skipped, because the NER module used for English texts (BBN IdentiFinder) is language specific and, if run on Slovenian, would output numerous false positives (spurious NE annotations). In the future, a language independent NE recognizer or a specific Slovenian module should be used; however, to our knowledge, no off-the-shelf Slovenian NE recognizer is readily available. Therefore, Slovenian texts imported into FrameNet Desktop currently do not include pre-tagged Named Entities.

The FrameNet Desktop import pipeline presupposes an existing installation of the TreeTagger (Schmid, 1994) for part-of-speech tagging and lemmatization. When importing a corpus, the path to a language-specific TreeTagger parameter file can be provided. Such parameter files are available for numerous languages from the TreeTagger website;[3] however, Slovenian is not among these. We therefore manually annotated a set of Slovenian texts, totaling approximately 14,000 running tokens, with a subset of the MULTEXT-East tagset (Erjavec, 2004) and trained a Slovenian model (parameter file). The "dictionary" file also necessary for training contains over 59,700 word forms.[4]

Once the corpus text has been fully processed by the import pipeline, the resulting XML is loaded into FrameNet Desktop by using the `FarinaImport` utility. The corpus files and their paragraphs show up as a tree structure in FrameNet Desktop when "Corpus mode" is selected as Tree Mode. Sentences of a paragraph can then be accessed individually for annotation, as illustrated in Figure 2 for the third sentence of the third paragraph from a speech by the Slovenian Prime Minister.

3.5 Annotation within FrameNet Desktop

Once a lexical unit has been attached to a frame, occurrences of its word forms in a corpus text are ready for annotation within FrameNet Desktop. For example, Figure 2 shows the annotation

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` [27 September, 2007].

[4] These resources are suitable for test purposes, but accuracy is not yet adequate for distribution.

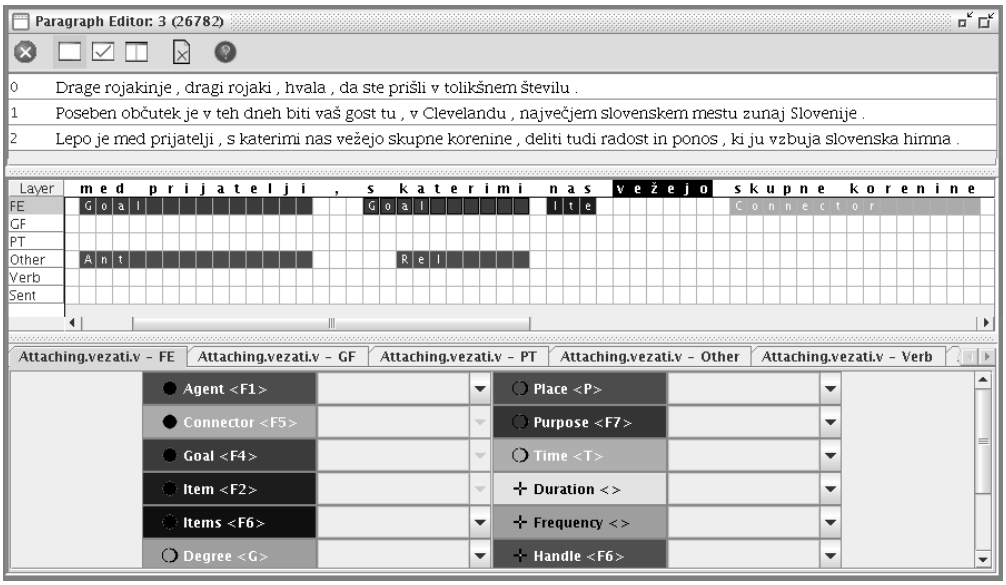


Figure 2: Annotation of Slovenian text within FrameNet Desktop.

of the target *vežejo* - '(they) tie', which is automatically recognized by FrameNet Desktop as a word form of the lexeme *vezati* - '(to) tie', and as an instance of the lexical unit *vezati.v* in the ATTACHING frame. Clicking on the corresponding pop-up text (`Attaching.vezati.v`) brings up several annotation layers relevant to this LU. On the frame element (FE) layer, words or word sequences within a sentence can be selected for annotation with one of the frame elements pertaining to the ATTACHING frame, displayed in a list at the bottom of the window. Other layers allow for the annotation of information on relative pronouns and antecedents, or for recording a metaphorical use of the target word, as in the example being annotated in Figure 2, as well as for the representation of other types of linguistic information.

A translation of the Slovenian sentence under consideration in Figure 2 is provided in Example (2).

(2) Lepo je med prijatelji, s katerimi
Nice it-is among friends, with whom
nas vežejo skupne korenine, deliti
us-ACC tie common roots, to-share
tudi radost in ponos, ki ju
also joy and pride that (them-DU)
vzbuja slovenska himna.
evokes Slovenian anthem.
'It is nice to share with friends, to whom we are tied by common roots, also the joy and pride evoked by the Slovenian anthem.'

3.6 Export from FrameNet Desktop

Corpora or documents annotated within FrameNet Desktop are exported with the `StaticReportGenerator` utility. This facilitates the creation of a wide variety of "reports". For the scenario under consideration, where we are interested in displaying annotated sentences as usage examples in an electronic dictionary, the most interesting report types are XML and HTML renderings of an annotated corpus. These represent, sentence by sentence as well as target word by target word, the frame element information attached to sentence constituents.

To illustrate, the XML representation of the annotation pertaining to the target lexical unit *vezati.v* in the ATTACHING frame, as shown in Figure 2, is given in Figure 3. The annotation with respect to *vezati.v* is subsumed under the `annotationSet` element, which provides as values of its attributes the names of the frame and lexical units, as well as their IDs in the MySQL database. `Target`, frame element (`FE`), and `Other` information attached to particular strings within the sentence is given as positions of their first and last character within the sentence string, as represented as values of `label` at-

```
<annotationSet ID="2334257" status="MANUAL" frameName=
   "Attaching" frameRef="197" luName="vezati.v"
   lexUnitRef="14542">
   <layers>
      <layer ID="11812320" name="Target">
         <labels>
            <label name="Target" ID="36094328"
                   start="40" end="45" />
         </labels>
      </layer>
      <layer ID="11812321" name="FE" rank="1">
         <labels>
    <label name="Connector" ID="36094330"
                   start="47" end="61" />
    <label name="Item" ID="36094332"
                   start="36" end="38" />
    <label name="Goal" ID="36094333"
                   start="8" end="21" />
    <label name="Goal" ID="36094334"
                   start="25" end="34" />
 </labels>
      </layer>
      [...]
      <layer ID="11812324" name="Other">
         <labels>
            <label name="Ant" ID="36094335"
                   start="8" end="21" />
    <label name="Rel" ID="36094336"
                   start="27" end="34" />
 </labels>
      </layer>
      <layer ID="11812325" name="Sent">
         <labels>
            <label name="Metaphor" ID="36094337"
                   itype="APos" />
         </labels>
      </layer>
      <layer ID="11812326" name="Verb" />
      </layers>
</annotationSet>
```

Figure 3: FrameNet XML representation of lexical unit annotation.

tributes. The XML report does not include information on the ATTACHING frame, such as its description or the default colors associated with each of its frame elements. Also, it does not explicitly state information on the lemma or lexeme underlying the annotated LU. These pieces of information, necessary to associate the sentence and its annotation with the relevant entry in the electronic dictionary, have to be retrieved from other types of FrameNet Desktop reports or from the MySQL database itself.

The HTML representation of annotated corpora consists of several files per target. One of them contains color markup of those annotated sentence constituents that are either target word or frame elements (see Figure 4). The information that the target has been annotated with respect to the ATTACHING frame is contained in a different HTML file, but none of the HTML files contains lexical unit information. As a result, reusing color information is easier from the HTML than from the XML representation of annotated sentences, but retrieving the correct lexeme for linking to a dictionary entry is even more difficult.

```
Lepo je <font style='color: #FFFFFF;background-color:
#006400'>med prijatelji</font> , <font style='color:
#FFFFFF;background-color: #006400'>s katerimi</font>
<font style='color: #FFFFFF;background-color:
#0000FF'>nas</font> <font style='color: #FFFFFF;
background-color: #000000'>

VEŽEJO

</font>
<font style='color: #FFFFFF;background-color:
#7CCD73'>skupne korenine</font> , deliti tudi radost
in ponos , ki ju vzbuja slovenska himna .
```

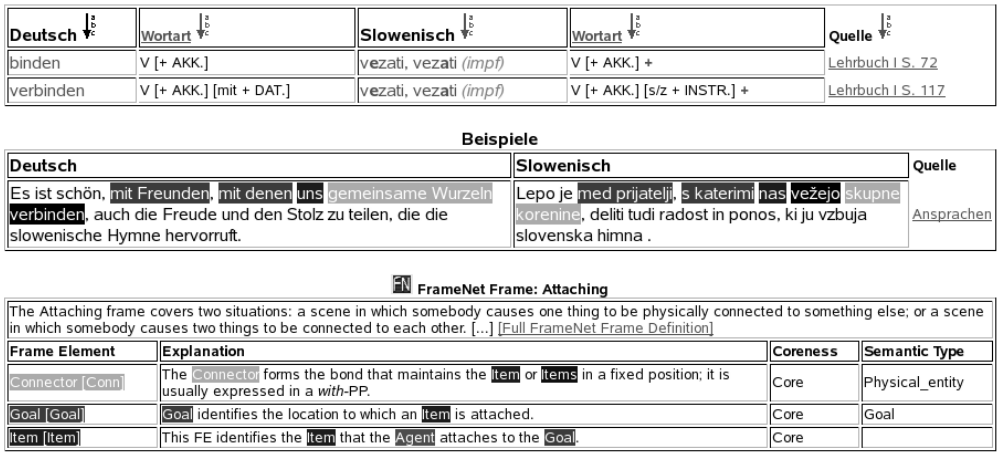
Figure 4: FrameNet HTML representation (selection) of lexical unit annotation.

We are currently studying the most elegant way to combine annotation set information with frame and lexeme information and to link the example to the correct dictionary entry. Once a method has been found, annotated example sentences could be shown within the dictionary as in Figure 5. The figure shows a bilingual setting, with an aligned sentence in German annotated for the corresponding German target word as well.

The user interface of the dictionary should display information useful to and accessible by humans, not necessarily experts in linguistics. Therefore, the frame-semantic markup simply represents target words and frame elements in different colors (see Figure 5). Definitions of the semantic frame and its frame elements (displayed below the examples) are kept to a minimum; a hyperlink leads to full frame-semantic information in the most current version of FrameNet.

4 Conclusion

We have presented FrameNet Desktop, an application for frame-semantic lexicon creation and corpus annotation. The recent remodeling of the FrameNet Desktop source code is intended to facilitate easy set-up of the tool, making it more accessible to new FrameNet-related projects. We used the new FrameNet Desktop version in a case study on a Slovenian FrameNet, aimed at data interchange with an electronic dictionary for this language. The actual set-up of a functional Slovenian FrameNet Desktop, as documented in this paper, including import of Slovenian lexicon data and corpus texts, but excluding



Deutsch	Wortart	Slowenisch	Wortart	Quelle
binden	V [+ AKK.]	vezati, vezati *(impf)*	V [+ AKK.] +	Lehrbuch I S. 72
verbinden	V [+ AKK.] [mit + DAT.]	vezati, vezati *(impf)*	V [+ AKK.] [s/z + INSTR.] +	Lehrbuch I S. 117

Beispiele

Deutsch	Slowenisch	Quelle
Es ist schön, mit Freunden, mit denen uns gemeinsame Wurzeln verbinden, auch die Freude und den Stolz zu teilen, die die slowenische Hymne hervorruft.	Lepo je med prijatelji, s katerimi nas vežejo skupne korenine, deliti tudi radost in ponos, ki ju vzbuja slovenska himna .	Ansprachen

FrameNet Frame: Attaching

The Attaching frame covers two situations: a scene in which somebody causes one thing to be physically connected to something else; or a scene in which somebody causes two things to be connected to each other. [...] [Full FrameNet Frame Definition]

Frame Element	Explanation	Coreness	Semantic Type
Connector [Conn]	The Connector forms the bond that maintains the Item or Items in a fixed position; it is usually expressed in a *with*-PP.	Core	Physical_entity
Goal [Goal]	Goal identifies the location to which an Item is attached.	Core	Goal
Item [Item]	This FE identifies the Item that the Agent attaches to the Goal.	Core	

Figure 5: Integration of frame-semantically annotated example sentences into dictionary.

training the Slovenian POS-tagger, took about two weeks. This is encouraging for any projects considering using the new FrameNet Desktop for their work.

The Slovenian Desktop does not yet offer the full functionality of its English counterpart, due to missing Named Entity Recognition and phrase type definitions. However, as the intended usage scenario focuses on collecting frame-semantic annotation in the form of frame element information, the current Slovenian FrameNet Desktop is already useful for preparing annotated example sentences that can be imported into an electronic dictionary.

5 Acknowledgments

This work was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD), granted to the first author.

References

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the FrameNet database. *International Journal of Lexicography*, 16(3):281–296.

Tomaž Erjavec. 2004. MULTEXT-East Morphosyntactic Specifications Version 3.0. `http://nl.ijs.si/ME/V3/msd/msd.pdf`, May.

Charles J. Fillmore, Miriam R.L Petruck, Josef Ruppenhofer, and Abby Wright. 2003. FrameNet in action: The case of attaching. *International Journal of Lexicography*, 16(3):297–332.

Primož Jakopin and Birte Lönneker. 2004. Query-driven dictionary enhancement. In *Proceedings of the Eleventh EURALEX International Congress*, pages 273–284, Lorient, France.

Birte Lönneker and Primož Jakopin. 2003. Contents and evaluation of the first German-Slovenian online dictionary. In *Proceedings of EACL 2003, Conference Companion*, pages 119–122, Budapest, Hungary.

Kyoko Hirose Ohara, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2004. The Japanese FrameNet project: An introduction. In *LREC 2004. The Fourth international conference on Language Resources and Evaluation. Proceedings of the Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"*, pages 9–11, Lisbon, Portugal.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. FrameNet II: Extended Theory and Practice. `http://framenet.icsi.berkeley.edu/book/book.pdf`, August.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Carlos Subirats and Hiroaki Sato. 2004. Spanish FrameNet and FrameSQL. In *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexically Annotated Corpora*, Lisbon, Portugal.

Carlos Subirats. Under review. Spanish FrameNet: A Frame Semantic analysis of the Spanish lexicon. In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications. (Under review)*. Mouton de Gruyter.

Interoperable Grammars

Michael Maxwell
Center for Advanced Study of Language/
University of Maryland
College Park, Maryland, USA
mmaxwell@casl.umd.edu

Anne David
Center for Advanced Study of Language/
University of Maryland
College Park, Maryland, USA
adavid@casl.umd.edu

Abstract

For languages with significant inflectional morphology, development of a morphological parser is often a prerequisite to further computational linguistic capabilities. We focus on two difficulties for this development: the short lifetime of software such as parsing engines, and the difficulty of porting grammars to new parsing engines. We describe a methodology we have developed to promote portability, using a formal declarative grammar written in XML, which we supplement with a traditional descriptive grammar. The two grammars are combined into a single document using Literate Programming. The formal grammar is designed to be independent of a particular parsing engine's programming language, thus helping solve the software lifetime and portability problems.

1 Grammar Development

After decades of widespread effort in computational linguistics, it is clear that progress has been made in areas ranging from the building computational lexical resources, to applications such as machine translation. While all this is beneficial, it is not without drawbacks: some resources which were developed at great expense, and which proved useful in the past, are no longer usable. This is perhaps nowhere more true than with grammars. Numerous computationally implemented grammars have been written, often at great expense; but nearly all of these grammars are tied to particular parsing engines, and their usefulness ends with the obsolescence of that parsing engine.

Recently, many computational linguists have turned from hand-crafted, labor intensive grammars to grammars automatically induced from annotated corpora, particularly in syntax. When the grammar is learned from a corpus, the obsolescence of a parsing engine may be a lesser issue, because when someone invents a better parsing tool, a grammar for that new tool can be induced from the same annotated corpus. This changes the issue from the obsolescence of grammars to the obsolescence of annotated corpora, and progress has been made in this area.

Automatic grammar induction has been more popular for syntax than for morphology. This is not to say that there has been no research into the learning of morphology; see for example Creutz and Lagus 2007, Goldsmith 2001, Goldsmith and Hu 2004, and the papers in Maxwell 2002. But research on morphology learning has not had the same impact that syntax learning has had. Accordingly, most wide coverage morphological parsers for languages with significant amounts of inflectional morphology are probably still built by hand; and barring a breakthrough, this seems likely to continue, at least for the near future.

Thus, for languages with significant inflectional morphology, a morphological parser[1] is a prerequisite to serious natural language processing. And to the degree that a language has complex morphology, the grammars for these parsers are difficult and time-consuming to build. One would therefore like to preserve this investment.

Unfortunately, the development of computer-processable morphological grammars is often tied to the programming language of a particular morphological parser, or to a general purpose computer programming language, such as Prolog or Haskel (see e.g. the papers at http://www.cs.chalmers.se/~markus/FM/index.html). If the particular morphological parser (or transducer) never became obsolescent, or if there were a standard descriptive lan-

[1] One commonly builds a morphological transducer, that is, a program which functions to both parse and generate inflected words. However, because it is more familiar, in this paper we will use the term 'parser.'

guage that all parsers used, this might not be problematic. But neither of these conditions is true. In the past 25 years, there have been at least half a dozen mutually incompatible morphological parsing languages, ranging from SIL's AMPLE (Weber Black and McConnel 1988) and PC-KIMMO (Antworth 1990) to Xerox PARC's xfst (Beesley and Karttunen 2003) Nor have developers of morphological parsing engines agreed upon a common language; two recent entries, the Stuttgart Finite State Transducer (http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html) and the OpenFst Library (http://www.openfst.org) provide still different programming languages. Some changes are motivated by enhanced capabilities, but others seem to be more an issue of style.

Two problems arise out of the mutual incompatibility of programming languages for different parsing programs: an Interoperability Problem, and a Half-Life Problem. The Interoperability Problem refers to the fact that a grammar written for one parsing engine cannot be used in another parser without re-writing; and for now, at least, that rewriting must be done by hand, since there are no automatic interpreters between parsing engine languages. A grammar written for the Xerox transducer, for example, will not run on the Stuttgart transducer without considerable modification.

The Half-Life Problem arises from the fact that software (in particular, parsers) becomes obsolete. While we are not aware of formal investigations, we estimate the average lifetime for language-based computational tools at five or ten years. In part, this is due to the (lack of) longevity of the underlying software.[2]

Software obsolescence can be postponed by the judicious choice of programming languages, avoiding OS-specific commands, the use of Open Source software, and the use of OS emulators. However, this can only prolong the life of a program, not extend it indefinitely. Few if any programs that were written in 1980 (twenty-seven years ago) still run on today's computers.

One might argue that software half-life is unimportant, since twenty years from now it may be possible to generate a morphological parser automatically from a corpus. Perhaps, but this remains to be seen. Meanwhile, the time and effort that go into writing grammars mandates that the grammars be usable long after the project is completed.

Another motivation for building longevity into parsing tools is that they constitute a description of (part of) the grammar of a language, in two senses. First, the grammar the parser uses constitutes a formal description of the language's morphology or syntax. Second, it can be used to analyze language texts, and—if it supports a generation mode—to produce paradigms. That is, a parser is an active description, not just a static one. But in their seminal paper, Bird and Simons 2003 point out that language data in computer-readable form can become unusable much more quickly than printed descriptions. In contrast, scholars of today can understand grammars and corpora penned thousands of years ago. Thus, while a parser constitutes a description of a language, it is—at present—an ephemeral description.

There is little doubt that future parsing engines will be improvements upon today's parsers. We are not suggesting that we need to build parsing engines which will continue to be used decades from now. Rather, we are suggesting that the language-specific information that goes into a parser—the grammar—should be written in such a way that it can be easily ported to future parsing engines. From this perspective, the Half-life Problem is really the Interoperability Problem in a different guise: interoperability between grammars written today, and tools which are yet to be built.

In summary, the problem is that while computational grammars have real worth, each parsing engine uses a different programming language. One might therefore conclude that there is no hope of providing a generic programming language for grammars. We claim that this conclusion is wrong; it is time now to think how we can write such grammars that will not only be interoperable with today's parsers, but with future parsing engines; and we provide a first cut at what such a programming language for morphology could look like.

One reason for optimism is the fact that morphological parsing tools now incorporate most of the capabilities that linguists have found necessary for morphology and phonology (albeit clumsily in some cases, e.g. morphosyntactic features), and that a morphological grammar written in a generic

[2] The first author, (Maxwell) was involved in a project in which two of the programming languages became defunct before the program was even complete; the cost of porting to alternative dialects of the programming languages was deemed prohibitive.

way can therefore be compiled into the programming language of current morphological parsers. At first glance, it might seem that this claim is simply wrong, because linguists have yet to come to agreement on the correct theory; like software, linguistic theories have a short half-life! Since the mid-1950s, there have been several generations of theories about what the phonological 'atoms' are, how many levels of structure are important, and how representations are translated between those levels. Atomic phonemes of the 1950s were replaced by distinctive feature matrices, which were in turn superseded by autosegmental representations. Nevertheless, these changes were primarily postulated to *explain* generalizations—generalizations which can be stated, if not explained so elegantly, with atomic phonemes. For example, a rule which spreads a feature of nasalization across vowels can be expressed as a rule that converts /a/ to /ã/, /o/ to /õ/, /u/ to /ũ/, etc. Phonological rules, and the natural classes used in those rules, can therefore be written in terms of atomic phonemes.

Similarly, while Optimality Theory (the current popular approach to phonology) holds that the phonetic form of words is determined by ranked constraints rather than rules, there is little if any empirical data that cannot be accounted for by a more traditional rule-based approach.[3]

In summary, while we may expect theories of morphology and phonology to continue to evolve, in practice it is quite feasible to do morphology and phonology using today's theories—or even yesterday's theories. Linguistics does not stand in the way of developing a sufficiently strong description language for morphology and phonology, and present-day finite state transducers are capable of implementing these descriptions.

The remainder of this paper sketches a design of a general language for writing morphological grammars, and a method for compiling grammars written in that generic language into the programming language of particular morphological parsers. We also describe how we supplement this formal grammar with a reference grammar of the morphology and phonology of a language. This may be seen as a way of commenting our code, but in fact we argue that it is much more, and that it constitutes a valuable effort in its own right.

[3] The under- and over-application of phonological rules in reduplication is a potential, if rare, counter-example.

2 Interoperability and Half-Life: Solution

We have embarked on a project to build morphological grammars and parsers of languages in a way that overcomes the Interoperability and Half-Life problems described in the previous section. The first grammar we have written is for the Bengali, or Bangla, language.

Our approach is to write a formal grammar in XML, using an XML schema to define the various grammatical structures needed to create—in conjunction with a suitable lexicon—a high coverage morphological parser. At present, this parser is being implemented in the Stuttgart Finite State Transducer (SFST). Since XML is not the native language of this parsing engine, we have written a 'compiler' to convert the XML-based grammar into code which SFST can use.

The following subsections describe our methodology in more detail.

2.1 Descriptive Grammar

During our investigation of Bangla, we were surprised to discover that no thorough and reliable descriptive grammar of modern colloquial Bangla exists, despite its having over 200 million native speakers. Instead, we relied on descriptions of Bangla morphology from half a dozen grammars, several journal articles, and a couple of dissertations. These sources were not always clear, nor did they always agree, and a few fine points of Bangla morphology were simply unexplained.

The difficulty we encountered in understanding grammatical descriptions, reconciling different grammatical accounts, and filling in gaps in coverage underline the fact that we could not have simply picked up an existing grammar and written our formal grammar from that. For languages which have any degree of inflectional complexity—and Bengali does, although there are languages with still more complicated morphologies—the complexities prevent such a simple approach. Instead, we began by writing a descriptive grammar, similar to reference grammars for other languages. It contains a chapter on the phonology and writing system of Bangla, plus chapters for the various parts of speech, describing the inflectional (and some derivational) affixes, and how the resulting inflected forms define the paradigms. The usage of these forms is also described with examples; it is not, however, a pedagogical grammar.

While the formal grammar described below was designed for interoperability, the reference grammar is much more than an add-on; rather, it is the means by which the formal grammar was written.

The following sections describe the formal grammar, and how we combined the descriptive and formal grammars into a unified whole.

2.2 Formal Grammar

In order to produce a morphological parser, one needs an unambiguous description of a language's morphology. Ambiguity is a fact about natural language, and one which plagues software specifications (Berry and Kamsties 2003). Building a parser from a descriptive grammar would be analogous to building traditional software from a specification. The danger is that our reference grammar, like the grammars we consulted, may be unclear or ambiguous, which would prevent its being used ten or a hundred years from now to build a new parser. We therefore need to supplement it with a grammar written in a formal language.

One approach would be to use the programming language of an existing parsing tool as that formal language. Amith and Maxwell (2005) propose using the xfst language (the language of one of the Xerox finite state tools, see Beesley and Karttunen 2003) for archival purposes. While this would meet our need for an unambiguous representation, it would fail to meet our goal of longevity: the Xerox tools will likely not be used in ten years, and there is no reason to think that the morphological parsing engines available then will use the same programming language, or that future grammar engineers will understand the xfst language.

Our formal grammar must therefore be not only unambiguous, but also—as far as possible—iconic and self-documenting. We decided to write our formal grammar in XML, and have developed an XML schema for this purpose.

The XML schema is based on a UML model developed by SIL (downloadable from http://fieldworks.sil.org/). This model allows for a rich set of morphological constructs:

- Item-and-arrangement affixes
- Item-and-process affixation
- Compounding and incorporation
- Paradigm classes, stem allomorphy classes
- A slot-and-template representation for inflectional affixation
- Morphosyntactic features structures
- Exception features
- Allomorphy constraints

Our schema allows for most of these constructs, with the exception for now of item-and-process morphology. We have supplemented the model with ordered phonological rules, which gives us a second mechanism for describing allomorphy.

Our XML schema is intended to "plug and play" with proposed standards for lexical databases, including the ISO draft Lexical Markup Framework (http://lirics.loria.fr/doc_pub/LMF_revision_14.pdf).

While we have built small test cases to exercise specific parts of the model and its schema, building a full-scale grammar allows us to test the schema in other ways. For example, our schema originally called for all regular expressions to be defined in one place, and called by reference (using XML *refid*s) in the various allomorph constraints and phonological rules where they are used. This worked well in small test cases, and it is computationally straightforward; it nevertheless turned out in our Bengali grammar to be too complex for linguists to maintain. As a result, we altered our schema to allow for regular expressions to be either called by reference, or defined where they are used. The former is used for regular expressions which are used often (such as the definition of consonants), while the latter is used for regular expressions that appear only once or twice.

2.3 Combining Descriptive and Formal Grammars

We have, then, both a descriptive and a formal grammar. We have argued elsewhere (Amith and Maxwell 2005a, 2005b) that neither is adequate by itself for long-term language description. We here summarize these claims.

First, we cannot presume that a linguist who was unfamiliar with our XML language could look at our formal grammar and easily deduce what it means. We therefore view our reference grammar as a supplement to the formal grammar. This is like commenting code, except that comments in traditional programs are intended for someone who is already conversant in the programming language,

whereas our reference grammar is intended to explain the meaning of the constructs to someone who does not already understand our XML grammar description language—a more ambitious goal.

On the other hand, since descriptive grammars are written in natural language, they are inherently ambiguous (as discussed above), and even vague. If a formal grammar could be combined with the descriptive grammar, we would have an antidote to this problem; assuming an appropriate syntax, a formal grammar is neither ambiguous nor vague.

The question then is whether the descriptive and formal grammars can be combined, allowing each to make up for the other's deficits. Such a combination would need the following components:

(1) A way to develop the grammars in parallel.
(2) A way to combine the grammars so that the description of each grammar topic is presented to the human reader along with the corresponding rules of the formal grammar.
(3) A way to extract the formal grammar for use by the parsing engine.

In fact, there already is a method that accomplishes (2) and (3); Literate Programming (henceforth LP), developed by Donald Knuth (Knuth 1984, 1992) for documenting computer programs. We have chosen the XML/ DocBook implementation of LP (Walsh and Muellner 1999; Walsh 2002), since XML provides advantages for long-term archiving (cf. the recommendations for the use of XML in Bird and Simons 2002). To this existing framework, we add a development methodology (described in David and Maxwell forthcoming), accomplishing point (1) above.

The result, we hope, is a mechanism that will allow another computational linguist—now or in the future—to pick up our grammar, understand what it means, and convert the formal grammar into the programming language of some other morphological parsing engine, either by writing an automatic converter, or by converting the grammar by hand.

As a reviewer pointed out, Literate Programming has not had a large impact on traditional software engineering. There are however two significant differences which give us reason to believe that LP can be more successful in computational linguistics. First, programmers are engineers, and they are notoriously resistant to writing documentation; and LP puts documentation first, so it is not surprising that software engineers are resistant. Linguistics, on the contrary, has a history of millennia of grammar documentation. Indeed, the problem for many linguists is exactly the opposite: describing grammars is natural, while writing formal language rules is unnatural. So if there is a problem in persuading linguists to do LP, it will be finding linguists who are willing to write the formal grammars (an issue addressed in the abovementioned methodology paper).

We suspect that another reason why software engineers are reluctant to spend the time doing LP is the burgeoning size of many computer programs. Programs which have been documented using LP are typically (if not always) small; but even utility programs, like the familiar command-line utilities of Unix, have become increasingly large and sophisticated. Grammars, on the other hand—at least morphological grammars—are comparatively simple, even for languages with complex morphologies. (Indeed, a grammar which seems too complex is often taken to be incorrect, or at the very least "missing a generalization.") Grammars, it seems to us, are just the right size for LP: not so small that they don't need documenting, and not so large that they cannot be documented. We are thus optimistic about the future of LP grammar writing.

2.4 Conversion to publishable grammar

We view our Bengali grammar, including both the descriptive and formal components, as a publishable work. Style sheets for DocBook of course exist already, giving us publication quality displays for our reference grammar. But while the formal grammar is understandable in its XML form, it is not "pretty" (as evident from the excerpt of our grammar in the appendix), nor does it bear an obvious resemblance to linguistic formalisms.

Fortunately, the flexibility of XML makes it possible to display a formal grammar using linguistic formalisms—for example, using style sheets to convert the XML structures for phonological rules into rules formatted in the way that linguists expect. The creation of the style sheets necessary to display and typeset our formal grammar is planned for next year, giving us the remaining piece needed for the Literate Programming. which Knuth referred to as 'weaving.'

2.5 Conversion to parser

The grammar is also intended to be used by a morphological parser. To build the parser, we first extract the formal grammar as an XML document

from the combined descriptive and formal grammar. This process, known in LP as 'tangling', is done by an XSLT program developed by Norman Walsh (available at http://docbook.sourceforge.net/release/litprog/current/fo/ldocbook.xsl).

The extracted XML formal grammar is then read by a Python program into an internal representation as objects, and output as the programming language of the target morphological parsing engine. A computer-readable lexicon must also be converted into the programming language of the parsing engine—a comparatively simple task.

Finally, the converted grammar and lexicon are read by the parsing engine, currently the Stuttgart Finite State Transducer, to produce the parser.

We expect any choice of parsing engine today to be superseded by more capable parsers. Targeting a different parsing engine will require rewriting only that part of the converter program that translates the program-internal representation into the target programming language (plus a separate converter for the lexicon). Alternatively, for relatively simple grammars it should be possible to translate an XML grammar into the target language by hand, a process aided by the side-by-side exposition of reference and formal grammars provided by the Literate Programming framework.

The analogy to the compilation of high-level programming languages is clear: while we compile our XML language into the high-level programming language of a morphological parsing engine, rather than into the machine language of a CPU, the goal is to make a program usable on a variety of platforms, both now and in the future.

Verifying that the conversion process works correctly with a parsing engine requires test data. Much of this test data can be automatically extracted from the descriptive grammar's paradigm tables and example sentences—another advantage of having both descriptive and formal grammars.

3 Previous work

We are not aware of previous work intended to produce grammars written in a formalism designed to be ported to different parsing engines. The closest work along these lines is perhaps DATR (Evans and Gazdar 1996), a formalism intended for lexical representation, incorporating a general mechanism for non-monotonic inheritance. It is possible to translate a DATR grammar into a morphological parser (see e.g. Colburn 1999). However, DATR is not an XML-based system. Perhaps more importantly, it uses a general purpose inheritance language, whereas our XML schema is a specialized language for morphology, allowing linguists to express linguistic constructs in linguistic terms.

SIL's recent FLEx program (http://www.sil.org/computing/fieldworks/flex/) is a database designed for linguistic field work. It incorporates a parser-independent representation of the morphological grammar; in fact, this is the source of the UML model that we used as the basis of our own XML schema. The SIL design anticipates that different parsers will be used in the future (FLEx currently uses SIL's XAMPLE parser), but the program is designed to support field linguists by providing a particular parser; FLEx was not designed with the goal of making it easy for computational linguists using other parsers to re-use grammars (Black and Simons 2006, Andrew Black, p.c.).

XLingPaper is an XML-based language developed by Andrew Black of SIL to write grammatical descriptions (see http://www.sil.org/~blacka/xlingpap/index.htm). XLingPaper allows for embedding interlinear text into documents, but it does not incorporate a formal grammar.

Some work on models and schemas for lexicons includes partial models of morphology, in particular the previously mentioned ISO draft Lexical Markup Framework (LMF). We are exploring adapting those parts of the 'intensional' and 'extensional' morphology specifications in LMF which overlap with our model. To some extent, LMF has been a moving target, and merging the two models will require further effort. Also, the ISO standard for feature structures (ISO 24610-1:2006) is used in our morphology model to represent morphosyntactic features.

There is also considerable work by computational linguists on defining models and schemas for lexicons and annotated text. We see our work as extending this to grammar development.

4 Conclusion

What is new about the project we describe is the development of an XML schema based on a model of morphology and phonology, and intended as a way of developing and documenting grammars so that they are easily ported to morphological parsing engines, both present and future.

While it is not a necessary part of modeling grammars for parser building, we believe that combining the formal XML-based grammar with a reference grammar intended to be read by humans provides increased portability. The combination serves as a better form of archival language documentation and description than either the reference grammar or the formal grammar by itself would.

Finally, we note that while our focus has been on morphological grammars, similar techniques—the development of a generic model, and the use of Literate Programming—could be applied to syntax. There is however perhaps less agreement on appropriate models among syntacticians than there is among morphologists, making this more of a hope than an immediately achievable goal.

References

Amith, Jonathan D., and Maxwell, Michael. 2005. Language Documentation: The Nahuatl Grammar. In Alexander Gelbuck (ed.) *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science. 474-485. Berlin: Springer.

Antworth, Evan L. 1990. *PC-KIMMO: a two-level processor for morphological analysis.* Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics.

Beesley, Kenneth R., and Karttunen, Lauri. 2003. *Finite State Morphology*: CSLI Studies in Computational Linguistics. Chicago: University of Chicago Press.

Berry, Daniel M., and Kamsties, Erik. 2003. "Ambiguity in Requirements Specification." In Julio Cesar Sampaio do Prado Leite and Jorge Horacio Doorn (eds.) *Perspectives on Software Requirements*. The Springer International Series in Engineering and Computer Science. Vol. 753. Berlin: Springer.

Bird, Steven, and Simons, Gary. 2002. Seven Dimensions of Portability for Language Documentation and Description. In *Proceedings of the Workshop on Portability Issues in Human Language Technologies, Third International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association.

Bird, Steven, and Simons, Gary. 2003. Seven dimensions of portability for language documentation and description. *Language* 79:557-582.

Black, H. Andrew, and Gary F. Simons. 2006. "The SIL FieldWorks Language Explorer Approach to Morphological Parsing." *Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society*, Austin, TX. http://www.sil.org/~simonsg/preprint/FLExParser%20Preprint.pdf

Butt, Myriam, King, Tracy Holloway, Niño, María-Eugenia, and Segond, Frédérique. 1999. *A Grammar Writer's Cookbook*: CSLI Lecture Notes, 95. Stanford, CA: CSLI Publications.

Colburn, Michael. 1999. "Enabling a Legacy Morphological Parser to use DATR-based Lexicons." Ph.D. dissertation, Colorado Technical University. http://ogea.org/Linguistics/Colburn2000.pdf.

Copestake, Ann, and Flickinger, Dan. 2000. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece.

Creutz, Mathias, and Lagus, Krista. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing* 4.

Cunningham, H., Tablan, V., Bontcheva, K., and Dimitrov, M. 2002. Language engineering tools for collaborative corpus annotation. http://citeseer.ist.psu.edu/734322.html.

David, Anne, and Michael Maxwell. Forthcoming. "Joint Grammar Development by Linguists and Computer Scientists." *Workshop on NLP for Less Privileged Languages, IJCNLP 2008*. Hyderabad.

Evans, R. and Gazdar, G. 1996. DATR: a language for lexical knowledge representation. *Computational Linguistics* 22: 167-216.

Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27:153-198.

Goldsmith, John , and Hu, Yu. 2004. From Signatures to Finite State Automata. *Midwest Computational Linguistics Colloquium*, Bloomington IN.

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27:97-111.

Knuth, Donald E. 1992. *Literate Programming*: CSLI Lecture Notes. Stanford: Center for the Study of Language and Information.

Ma, Xiaoyi, Lee, Haejoong, Bird, Steven, and Maeda, Kazuaki. 2002. Models and Tools for Collaborative Annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association.

Maxwell, Michael B. 2002. *Proceedings of the Workshop on Morphological and Phonological Learning*. New Brunswick, NJ: ACL.

Nirenburg, Sergei, Biatov, Konstantin, Farwell, David, Helmreich, Stephen, McShane, Marjorie, Ponsford, Dan, Raskin, Victor, and Sheremetyeva, Svetlana. 1999. Toward Descriptive Computational Linguistics. http://crl.nmsu.edu/expedition/publications/boas-acl99.pdf.

Oepen, Stephan, Flickinger, Dan, Tsujii, Jun-ichi, and Uszkoreit, Hans. 2001. *Collaborative Language Engineering: A Case Study in Efficient Grammar-Based Processing*: CSLI Lecture Notes, 118. Chicago: University of Chicago Press.

Oflazer, Kemal, Nirenburg, Sergei, and McShane, Marjorie. 2001. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics* 27:59-85.

Walsh, Norman, and Muellner, Leonard. 1999. *DocBook: The Definitive Guide*. Sebastopol, California: O'Reilly & Associates, Inc.

Walsh, Norman. 2002. Literate Programming in XML. *XML 2002*, Baltimore, MD.

Weber, David; Andrew Black; and Stephen R. McConnel. 1988. *AMPLE: A tool for exploring morphology*. Occasional Publications in Academic Computing no. 12. Dallas: Summer Institute of Linguistics.

Appendix: Sample Grammar Excerpt

3.2. Future Tense

The future tense is used to express:

- a future state or action
- propriety or ability [*etc.*]

…

Table 6.2. FutureTense Verb Forms

Person	Suffix	(C)VC-	(C)aC-	(C)V-	(C)a-	(C)V(i)-	Causative	3-অক্ষর
		শোনা /ʃon-a/ *to hear*	থাকা /thak-a/ *to stay*	হওয়া /hɔ-oya/ *to become*	খাওয়া /kha-oya/ *to eat*	চাওয়া /ca-oya/ *to want*	শেখানো /ʃekha-no/ *to teach*	কাম্‌ড়ানো /kamṛa-no/ *to bite*
1st	-বো /-bo/	শুন্‌বো /ʃun-bo/	থাক্‌বো /thak-bo/	হব /hɔ-bo/	খাব /kha-bo/	চাইব /cai-bo/	শেখাব /ʃekha-bo/	কাম্‌ড়াব /kamṛa-bo/

[*Additional rows omitted to save space*]

The formal grammar's listing of future tense suffixes appears below.

```
<Mo:InflectionalAffix gloss="-1Fut" id="af1Fut">
   <!--The two "allomorphs" are really allographs-->
   <Mo:Allomorph form="বো">
      <!--Spelled 'bo'; usually (not always) after a C-stem -->
   </Mo:Allomorph>
   <Mo:Allomorph form="ব">
      <!--Spelled 'b'; usually (not always) after a vowel stem -->
   </Mo:Allomorph>
   <Mo:inflectionFeatures>
     <Fs:f name="Tense"><Fs:symbol value="Future"/></Fs:f>
     <Fs:f name="Mood"><Fs:symbol value="Indicative"/></Fs:f>
     <Fs:f name="Person"><Fs:symbol value="1"/></Fs:f>
   </Mo:inflectionFeatures>
/Mo:InflectionalAffix>

<!-- Etc. for the remaining future tense suffixes -->
```

A Machine Learning Approach to Building Aligned Wordnets

Gerard de Melo
Max Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken, Germany
demelo@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Campus E1 4
66123 Saarbrücken, Germany
weikum@mpi-inf.mpg.de

Abstract

WordNet is a lexical database describing English words and their senses. We propose a method for automatically producing similar resources for new languages by taking advantage of the original WordNet in conjunction with translation dictionaries. A small set of training mappings is used to learn a model for predicting associations between terms and senses. The associations are represented using a variety of scores that take into account structural properties as well as semantic relatedness and corpus frequency information. For evaluation, we created a German-language wordnet, and the data indicate a significantly better coverage and higher precision than previous heuristics. The resulting resources provide not only valuable information for monolingual NLP tasks but also enable a high degree of cross-lingual interoperability.

1 Introduction

Princeton WordNet (Fellbaum, 1998) is a well-known lexical resource that provides information about how words and word senses in the English language are linked. It lists the senses that a word can assume and provides structural information about how such senses are related, e.g. via the hypernymy relation that holds when one term is a generalization of another term, e.g. "`publication`" is a hyponym of "`journal`". The original WordNet for the English language later inspired endeavours to create similarly structured databases ("wordnets") for other languages, e.g. in the context of the EuroWordNet EU project (Vossen, 1998), the BalkaNet project, as well as under the auspices of the Global WordNet Association. Nevertheless, we contend that despite several decades of work on such resources, there is still a great need for additional research into more efficient means of producing them. Consider, for instance, that there are about 7,000 living languages, but only around 40 for which wordnet versions have been created, many indeed still in a preliminary stage with very low coverage, and less than a handful of languages with wordnet versions that are freely downloadable from the Internet. Furthermore, several existing wordnets unfortunately form completely independent networks that are not connected in any way to other wordnets.

In order to complement the existing manually compiled wordnets, we thus propose a new approach to building wordnets that trades off accuracy for a much faster compilation process, and hence frequently leads to more terms being covered than in existing wordnets. Our approach is based on learning classifications, and therefore is completely automatic once an initial set of training mappings is provided. Certainly, the resulting wordnets will not have the same level of accuracy as resources carefully constructed by expert lexicographers, however they can 1) serve as a valuable starting point for creating more accurate ones, and 2) be used immediately in many natural language processing tasks where coverage is more important than perfect accuracy. The fact that the wordnets are aligned with the Princeton WordNet greatly facilitates interoperabil-

ity with existing wordnets (e.g. English-language glosses are available) as well as with additional resources such as topical domain labels (Bentivogli et al., 2004) or mappings to ontologies (Niles and Pease, 2003; Suchanek et al., 2007). This additional information is an enormous benefit in practical applications.

The remainder of this paper is organized as follows. Based on a brief introduction to classification learning in Section 2, we present our wordnet building approach in Sections 3 and 4. This approach is then evaluated in Section 5, while Section 6 compares it to other existing work in this field. Finally, concluding remarks are provided in Section 7.

2 Preliminaries

A *classification* is an assignment of class labels $y \in \mathcal{Y}$ to objects $x \in \mathcal{X}$ in the form of a function $\widehat{f} : \mathcal{X} \times \mathcal{Y} \longrightarrow \{\top, \bot\}$ where $\top$ indicates the object being assigned the class label, and $\bot$ indicates the contrary. We consider only binary problems, where $\mathcal{Y} = \{C, \overline{C}\}$ for some class C and its complement $\overline{C}$. Learning a classification then consists in finding a function f that approximates a true classification $\widehat{f}$ with low approximation error, given a set of correctly labelled training examples $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Provided that the objects $x \in \mathcal{X}$ are represented in a suitable manner, usually as numerical *feature vectors* $\mathbf{x}$ in an m-dimensional Euclidean feature space $\mathbb{R}^m$, one of several learning algorithms can be employed to learn a classification. Support vector machines constitute a class of algorithms based on the idea of computing a decision hyperplane $\mathbf{w}^T\phi(\mathbf{x})+b = 0$ that maximizes the margin between positive and negative training instances in the feature space (Vapnik, 1995). Such maximum-margin hyperplanes tend to entail lower generalization errors than other separation surfaces, and the task of finding them leads to a quadratic optimization problem. Additional slack variables can be included for a soft margin solution that is able to deal with training data that cannot be separated cleanly (Cortes and Vapnik, 1995). The decision surface can then be computed using Lagrange multipliers and decomposing techniques such as sequential minimal optimization (Platt, 1999).

3 Building Wordnets by Learning Classifications

In order to build wordnets automatically, we suggest the following approach. Let L_{N} denote the language for which a wordnet is to be constructed, and L_0 denote the language of an existing wordnet that serves as a template for the new one, in our case the English language due to our choice of Princeton WordNet as the template. The existing wordnet immediately provides a sense inventory as well as information about the links between the senses, though certain relations need to be interpreted as generic relatedness links between senses (e.g. the derivation relation), or are completely excluded from being adopted (e.g. region domains). The most significant missing ingredient at this point are the links from terms in L_{N} to their respective senses. This is tackled by means of translation dictionaries, however with the constraint of relying on a minimal amount of information specific to L_{N} so that the approach remains generalizable to as many languages as possible. The dictionary is thus conceived as offering a simple $n : m$-mapping between terms in L_0 and terms in L_{N}, with optional part of speech information, as in the following excerpt:

```
{n}   Schulabbrecher  -  dropout
      ...                ...
{n}   Schulklasse     -  class
{n}   Schulklasse     -  form
      ...                ...
      schulmäßig      -  scholastic
{adv} schulmäßig      -  scholastically
```

Given a translation from a term t from L_{N} to a term e from L_0, one may assume that there is very likely some semantic overlap between t and e, so some sense of e is likely to also be a sense of t. We thus proceed as follows: for each term t from L_{N}, retrieve the set of translations $\phi(t)$. For each L_0-translation e in such a $\phi(t)$, retrieve the set of senses $\sigma(e)$ from our existing wordnet, e.g. for the German term "`Schulklasse`" the senses of the translations "`class`" and "`form`" would be considered. Our goal is now to determine for each sense $s \in \bigcup_{e \in \phi(t)} \sigma(e)$ whether s is also an appropriate sense of t. This is undoubtedly a very difficult task, as the dictio-

naries provide only limited information that could aid in determining which of the often many different senses listed by WordNet apply, e.g. 9 senses for the word "class" and 23 senses for "form". In our approach, the problem is construed as a binary classification problem. A real-valued feature vector $\mathbf{x}$ is created for each pair (t, s) of a term t from L_{N} and a relevant candidate sense s from the wordnet for L_0. For example, if t represents "Schulklasse", then s could be one of the senses of "class". In order to create the feature vectors, a variety of different fitness scores x_i are used as features and combined as components of numeric vectors $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X} = \mathbb{R}^m$.

Based on a small set of manually established labels for such (t, s)-pairs, we create the corresponding training set of feature vectors and derive a classification model that can be used to make predictions for any other (t, s)-pair. Assuming that the model provides confidence scores $c_{t,s}$ for (t, s)-pairs, we apply one of the following rules for every L_{N}-term t from the translation dictionary, combined with each of its possible candidate senses as defined above:

a) accept as a weighted connection with weight $c_{t,s}$ whenever $c_{t,s} > 0$, or
b) accept as an unweighted connection whenever $c_{t,s} \geq \alpha_1$ or $\forall s' \neq s : c_{t,s} > c_{t,s'}$ (for two pre-defined constants α_1 and $\alpha_2 \leq \alpha_1$).

The first rule results in a weighted statistical wordnet for L_{N}, whereas the second one yields a conventional unweighted wordnet. Finally, new senses may be introduced manually to cover terms for which no candidate senses were found.

This approach has several advantages compared to the previous work in this field (cf. Section 6). First of all, the previous automatic approaches were based on hard acceptance criteria - either a (t, s)-pair satisfies a criterion or not. Many attributes of word senses do not lend themselves easily to such an antagonistic view, e.g. sense relatedness measures produce numeric scores, and thus can be better accommodated in a model that uses real-valued feature vectors. Furthermore, while Atserias et al. (1997) investigate combinations of two heuristics to arrive at a greater accuracy, a classification learning approach can take into account suitable combinations of even more heuristics, indeed arbitrary linear (or even non-linear) combinations of feature values.

4 Feature Computation

Following the description of the overall procedure, we will now go into more detail on how the feature values x_i are computed before being combined to create the feature vector for a given (t, s)-pair.

4.1 Lexical Category Compatibility

Unlike previous work, our study considers all lexical categories (parts of speech) covered by the existing wordnet rather than just nouns. This immediately leads to the problem that the number of candidate senses greatly increases, and we need to come up with some means of preventing a noun from being mapped to a verb sense in WordNet, for instance.

Our solution rests on two pillars. Obviously, whenever the translation dictionary explicitly provides lexical category information, one can simply use hard-coded compatibility indicators, e.g. we give any German adjective a compatibility value of 0.0 with English noun senses, but 1.0 with English adjective as well as adverb senses.

In light of the fact that such information may not always be available, we resort to additional heuristics when such explicit information is not available, thereby ensuring that our approach remains applicable to a broad range of different scenarios. For each lexical category, a C4.5 decision tree is used to estimate the compatibility based on superficial attributes of the terms such as suffixes and capitalization. Growing the trees does not require any manually created training data, because we can leverage terms where all candidate senses share the same lexical category. The features employed are given in the following list. Note that since the terms in L_{N} can be multi-word expressions, much of this information is captured separately for the first and last word of any candidate expression.

- prefixes of the first and last word up to a length of 10, e.g. for the German verb "einschulen", "e", "ei", "ein", etc. would be considered
- suffixes of the first and last word up to a length of 10 (without case conversion), e.g. "n", "en", "len", etc. for "einschulen".

- capitalization of the first and last word (Boolean features for no capitalization, capitalized first character, and all characters capitalized)
- term length

The decision trees were pruned to have confidence levels of at least 0.25 with at least 2 instances per leaf. The confidence estimations from the leaves can then be used to determine a lexical category compatibility score as a feature in the feature vector. For languages where the predictions are too unreliable, we may instead use a constant value of 0.5.

4.2 Sense Weighting Functions

Several features that will be described later on depend on some kind of assessment of the importance of senses s with respect to the particular L_{N}-term t under consideration. We consider the following weighting functions $\gamma(t, s)$:

- $\gamma_1(t, s) = 1$ is used for unweighted features
- $\gamma_{\mathrm{lc}}(t, s)$ represents an estimation of the lexical category compatibility between t and s, as described earlier
- $\gamma_{\mathrm{r}}(t, s)$ considers the ranks of the senses as listed by WordNet for the translations of t, as these are indicators for the importance of a sense. It is computed as follows:

$$\gamma_{\mathrm{r}}(t, s) = \gamma_{\mathrm{lc}}(t, s) \left[\sum_{e \in \phi(t)} \frac{1}{r(e, s)} \right]$$

  where $r(e, s)$ yields 1 if s is the highest-ranked sense for e, 2 for the second sense, and so on.
- $\gamma_{\mathrm{f}}(t, s)$ considers the corpus frequency information provided with WordNet:

$$\gamma_{\mathrm{f}}(t, s) = \gamma_{\mathrm{lc}}(t, s) \left[\sum_{e \in \phi(t)} \frac{f(e, s)}{\sum_{s' \in \sigma(e)} \lambda_{s,s'} f(e, s')} \right]$$

  where $f(e, s)$ returns the number of occurrences of term e with sense s in the corpus, and $\lambda_{s,s'}$ is 1 if the lexical category of s and s' match, and 0 otherwise.

4.3 Semantic Relatedness Measures

Apart from weighting functions, our approach is fundamentally based on measures of semantic relatedness between senses, e.g. the single sense of "`schoolhouse`" is related to the educational institution sense of "`school`", but not to the sense of "`school`" that refers to groups of fish. Before going into details of how semantic relatedness contributes to many of our fitness scores, we shall first introduce several relatedness estimation heuristics.

- $\mathrm{sim}_{\mathrm{id}}(s_1, s_2)$ is simply the trivial identity indicator function, i.e. yields 1 if $s_1 = s_2$, and 0 otherwise.

$$\mathrm{sim}_{\mathrm{id}}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 0 & \text{otherwise} \end{cases}$$

- $\mathrm{sim}_{\mathrm{f}}(s_1, s_2)$ considers not only whether two senses are identical but also takes into account senses that stand in a parent-child or sibling relationship in terms of the hypernym hierarchy.

$$\mathrm{sim}_{\mathrm{f}}(s_1, s_2) = \begin{cases} 1 & s_1 = s_2 \\ 0.8 & \text{hypernymy/hyponymy} \\ 0.7 & \text{siblings, no hypernymy} \\ 0 & \text{otherwise} \end{cases}$$

- $\mathrm{sim}_{\mathrm{n}}(s_1, s_2)$ considers the graph neighbourhood and acknowledges relations other than hypernymy/hyponymy as well as transitive connections (e.g. a holonym of a hypernym). For a given path in the graph, we may compute an inverse distance score multiplicatively from relation-specific edge weights (e.g. 0.8 for hypernymy, 0.7 for holonymy). The relatedness score is then defined as the maximum score for all paths between s_1 and s_2 if this maximum is above or equal a pre-defined threshold $\alpha_n = 0.35$, and 0 otherwise. It can be obtained efficiently using a Dijkstra-like algorithm (de Melo and Siersdorfer, 2007).
- $\mathrm{sim}_{\mathrm{c}}(s_1, s_2)$ uses the cosine similarity of context strings for senses, which are constructed by concatenating glosses and lexicalizations of the sense itself with those of senses directly related via hyponymy, holonymy, derivation, or instance relations, as well as with those of 2

levels of hypernyms. The terms are stemmed using Porter's stemmer, and feature vectors $\mathbf{v}_1$, $\mathbf{v}_2$ with TF-IDF values are created based on the bag-of-words vector space model. The score is then computed as the cosine of the angle between the vectors, i.e. as $\mathbf{v}_1^T \mathbf{v}_2 (||\mathbf{v}_1|| \, ||\mathbf{v}_2||)^{-1}$.

- $\mathrm{sim}_\mathrm{m}(s_1, s_2)$, finally, is simply defined as $\max\{\mathrm{sim}_\mathrm{f}(s_1, s_2), \mathrm{sim}_\mathrm{n}(s_1, s_2), \mathrm{sim}_\mathrm{c}(s_1, s_2)\}$, and hence combines the power of sim_f, sim_n, and sim_c, which is particularly valuable due to the fact that sim_n and sim_c are based on very different characteristics of the senses.

4.4 Semantic Overlap Features

One important way of making use of the semantic relatedness measures is to exploit that a mapping should more likely be accepted when a term t has multiple English translations e, and the candidate sense s under consideration is somewhat pertinent to multiple of them. For instance, the German "`Schulklasse`" has the terms "`class`" and "`form`" as translations. While "`form`" can not only refer to a body of students who are taught together but also e.g. to a tax form, only the former of these two senses overlaps semantically with the senses of "`class`".

Given a term t and a candidate sense s, we integrate scores of the following form into the respective feature vector:

$$\sum_{e \in \phi(t)} \max_{s' \in \sigma(e)} \gamma(t, s') \, \mathrm{sim}(s, s') \tag{1}$$

$$\sum_{e \in \phi(t)} \frac{\sum\limits_{s' \in \sigma(e)} \gamma(t, s') \, \mathrm{sim}(s, s')}{\sum\limits_{s' \in \sigma(e)} \gamma(t, s')} \tag{2}$$

where $\mathrm{sim}(s_1, s_2)$ represents a semantic relatedness measure and the $\gamma(t, s)$ function provides weights as described earlier. The simple identity relatedness function sim_id and the constant weighting function $\gamma_1(t, s) = 1$ make Equation 1 yield a simple count of how many English terms are mapped to the sense, reminiscent e.g. of the equivalent word matching of Okumura and Hovy (1994) (cf. Section 6). By using the above formulae to produce a large number of feature values with all combinations of weighting functions and relatedness measures mentioned in Sections 4.2 and 4.3, we are able to account for cases where the terms are related but do not share senses.

4.5 Polysemy-Based Scores

Another set of features are based on the polysemy of the L_0-translations, i.e. on the idea that a mapping is more likely correct whenever there are few alternative senses to choose from. Akin to the monosemy heuristic of Okumura et al. (see Section 6), we can consider for instance the German "`Schulleiter`" with its translation "`headmaster`", which in turn only has one single sense listed in WordNet, so it is rather safe to accept this sense also for the German term. More generally, given a term t and a sense s, several scores can be computed as

$$\left(1 + \sum_{s' \in C} \gamma(t, s')(1 - \mathrm{sim}(s, s'))\right)^{-1} \tag{3}$$

where $\gamma(t, s)$ is a weighting function and $C = \bigcup\limits_{e \in \phi(t)} \sigma(e)$ stands for the complete candidate set. Another set of scores is computed as

$$\sum_{e \in \phi(t)} \frac{\mathbf{1}_{\sigma(e)}(s)}{1 + \sum\limits_{s' \in \sigma(e)} \gamma(t, s')(1 - \mathrm{sim}(s, s'))} \tag{4}$$

where $\mathbf{1}_{\sigma(e)}(s)$ is the indicator function for $\sigma(e)$, and therefore yields 1 if $s \in \sigma(e)$ and 0 otherwise.

Again, we can use $\mathrm{sim}_\mathrm{id}(s_1, s_2)$ and $\gamma_1(t, s)$ to illustrate the simplest case: Equation 3 then yields the reciprocal of the total number of candidate senses and in Equation 4 the denominator of each addend becomes 1 whenever the respective term e is monosemous according to WordNet. More advanced scores are computed by

- using Equations 3, 4 with $\gamma_1(t, s)$, combined with either sim_f, sim_c, sim_n, or sim_m, and
- using Equation 4 with $\mathrm{sim}_\mathrm{id}(s_1, s_2)$ and one of the weighting functions $\gamma_\mathrm{lc}(t, s)$, $\gamma_\mathrm{r}(t, s)$, or $\gamma_\mathrm{f}(t, s)$.

4.6 Additional Features

We further consider a number of other, less essential features, including the following:

- scores based on the number of translations

$$\left(\sum_{e \in \phi(t)} \lambda(t, e) \right)^{-1}$$

as well as the ratio

$$\frac{\sum_{e \in \phi(t)} \lambda_{\text{wn}}(t, e)}{\sum_{e \in \phi(t)} \lambda_{\text{id}}(t, e)} = \frac{\sum_{e \in \phi(t)} \lambda_{\text{wn}}(t, e)}{|\phi(t)|}$$

where $\lambda(t, e)$ is a translation weighting function that can be either $\lambda_{\text{id}}(t, e) = 1$ or $\lambda_{\text{wn}}(t, e)$, which is 1 if $\sigma(e) \neq \emptyset$, and 0 otherwise.

- a score based on back-translations

$$\sum_{e \in \phi(t)} \frac{\mathbf{1}_{\sigma(e)}(s)}{|\phi^{-1}(e)|}$$

where $\phi^{-1}(e)$ is defined as $\{t \mid e \in \phi(t)\}$.

- the number of lexicalizations of the candidate sense, i.e. $\left|\sigma^{-1}(s)\right|$, where $\sigma^{-1}(s)$ is defined as $\{e \mid s \in \sigma(e)\}$.
- the ratio of sense lexicalizations that are translations of t, i.e.

$$\frac{\sum_{e \in \sigma^{-1}(s)} \lambda_{\text{tr}}(t, e)}{|\sigma^{-1}(s)|}$$

where $\sigma^{-1}(s)$ is defined as above, and $\lambda_{\text{tr}}(t, e)$ yields 1 if $e \in \phi(t)$ and 0 otherwise.

- indicator values that express whether the candidate sense s is a noun, verb, adjective, or adverb sense, respectively.

5 Experimental Evaluation

While our approach is applicable to virtually any language, we evaluated it by generating a German wordnet based on the Ding German-English dictionary (Richter, 2007), a large and fairly reliable digital translation dictionary with around 216,000 entries, but not much additional information apart from (optional) part of speech tags. Princeton WordNet 3.0, which covers around 155,000 English terms and around 118,000 senses, served as the existing template for the new wordnet.

Table 1: Comparison with existing methods

	precision	recall
First Sense Heuristic	40.36%	67.46%
Rigau & Agirre	48.97%	63.58%
Atserias et al.[1]	75.00%	35.82%
Benítez et al.	73.14%	38.21%
Our Approach	81.11%	65.37%

[1]: excluding criteria based on additional background knowledge (see text)

We manually evaluated 1,834 candidate mappings for 350 randomly selected German terms from the dictionary for use as training data (407 mappings, i.e. 22%, were positive). To create a test set with both positive and negative examples, the same was repeated with another 1,624 candidate mappings for 350 further randomly selected terms. Based on this training data, the LIBSVM implementation (Chang and Lin, 2001) of support vector machine learning was used to derive a linear kernel model and additionally also estimate posterior class probabilities for the (t, s)-pairs using a variant of Platt's method (Lin et al., 2007). The thresholds $\alpha_1 = 0.5$ and $\alpha_2 = 0.45$ were applied on these estimates as described in Section 3 to generate a German wordnet.

Our technique is compared to four alternative approaches. We study the first sense heuristic, which involves simply accepting the first sense listed by WordNet for any English term, and is frequently cited as more successful than many other heuristics in word sense disambiguation tasks because the rank reflects the corpus frequency and importance of a sense. We also evaluate existing automatic approaches presented in Section 6. From the study by Atserias et al. (1997), we consider the monosemy 1-4, variant, as well as the combined brother and polysemy 1/2 criteria. The CD criteria and the field criterion were not applied because their implementation in the original study is mainly based on additional lexical information for the Spanish language apart from the list of translations. The standard classification evaluation measures of precision and recall were used. Given a test set, the precision is defined as $\frac{P_T}{P_T+P_F}$, and the recall is defined as $\frac{P_T}{P_T+N_F}$, where P_T, P_F, N_F are the sets of true positives, false positives, and false negatives, respectively. The results,

Table 2: Alternative confidence thresholds

α_1	α_2	precision	recall
0.90	0.80	94.21%	34.03%
0.90	0.60	91.50%	41.79%
0.70	0.60	87.50%	52.24%
0.60	0.50	83.90%	59.10%
0.50	0.45	81.11%	65.37%
0.40	0.35	73.64%	72.54%
0.35	0.25	70.53%	80.00%
0.30	0.25	67.32%	82.39%
0.20	0.15	55.93%	90.15%
0.10	0.05	40.41%	94.93%

Table 3: Coverage Statistics

	sense mappings	terms	lexicalized senses
nouns	53,146	35,089	28,007
verbs	13,875	5,908	6,304
adjectives	21,799	13,772	9,949
adverbs	4,243	2,992	2,593
total	93,063	55,522	46,853

presented in Table 1, demonstrate that our learning-based approach outperforms the existing approaches both in terms of precision as well as in terms of recall. While two previous heuristics arrive at similarly high levels of recall, this occurs at the expense of very low precision scores. By adjusting the α_1, α_2 confidence thresholds, our method can be made to produce recall scores well above 90% at such levels of precision. Table 2 provides a sample of results obtained using alternative thresholds.

In addition to the recall scores in Table 1, which are based on the test set, we also provide the absolute number of terms covered by the resulting German wordnet in Table 3. The figures are below the current size of GermaNet 5.0 (Kunze and Lemnitzer, 2002), but larger by an order of magnitude than many other manually compiled wordnets.

6 Related Work

Although no other studies have considered building new wordnets by classifying real-valued feature vectors, there has been prior work on heuristics for linking dictionaries to WordNet. Knight (1993) created an ontology for machine translation by linking entries in Longman's Dictionary of Contemporary English to WordNet, taking into account gloss definitions as well as the semantic hierarchy information present in the dictionary, though unfortunately not available in our setting. Okumura and Hovy (1994) used a Japanese-English dictionary to link a Japanese lexicon to this ontology, based on several heuristics, most importantly monosemy, i.e. considering when the ontology lists only one candidate concept for an English translation, and equivalent word matches, i.e. accepting the concepts shared by multiple translations of a word.

Rigau and Agirre (1995) presented a preliminary study on mapping Spanish nouns to WordNet senses by looking up the translations of the Spanish noun, and then checking whether the senses of those translations satisfy certain criteria. Atserias et al. (1997) proposed additional heuristics for generating a preliminary noun-only version of the Spanish WordNet that later were adapted for producing preliminary noun-only Catalan and Hungarian wordnets (Benitez et al., 1998; Miháltz and Prószéky, 2004).

Pianta et al. (2002) used similar techniques in conjunction with a cosine similarity-based heuristic to create rankings of the most likely candidate senses that were then presented to human lexicographers for selection. This methodology was used to create MultiWordNet Italian and later also the Hebrew WordNet. Sathapornrungkij and Pluempitiwiriyawej (2005) used criteria proposed by Atserias et al. (1997), and then performed a regression analysis in order to reduce the number of accepted mappings and thus increase the accuracy. Since they merely relied on 12 binary criteria rather than numeric scores, they were unable to obtain a higher recall by applying their model to other term-sense pairs not fulfilling one of the chosen criteria.

7 Conclusions

We have shown that wordnets can be built automatically if we are willing to accept a certain percentage of imprecise mappings. Our approach based on learning from a number of numeric scores leads to a better coverage than the hard criteria proposed in previous studies, while simultaneously also allowing for a higher level of accuracy. It is fair to as-

sume that the method presented scales well to new languages, because care was taken to require just a minimal amount of information specific to L_{N}. The resulting resources greatly facilitate interoperability, as they are aligned to the original Princeton WordNet, and thus also to other resources that are similarly aligned.

In the future we would like to investigate techniques for extending the coverage of such automatically generated wordnets to senses not covered by the existing wordnet. It is well-known that for a variety of tasks one can benefit from the information stored in lexical resources, e.g. for word sense disambiguation, for query expansion in information retrieval, especially in image and multimedia retrieval, and for cross-lingual applications. We will soon provide a more detailed analysis of the quality of automatically generated wordnets, also studying in detail their suitability for use in monolingual as well as cross-lingual natural language processing tasks.

References

Jordi Atserias, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual WordNets. In *Proc. Intl. Conf. on Recent Advances in NLP 1997*, pages 143–149.

Laura Benitez, Sergi Cervell, Gerard Escudero, Monica Lopez, German Rigau, and Mariona Taulé. 1998. Methods and tools for building the Catalan WordNet. In *Proc. ELRA Workshop on Language Res. for Europ. Minority Lang., 1st Intl. Conf. on Language Resources & Evaluation*.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising the WordNet domains hierarchy. In *COLING 2004 Multiling. Ling. Resources*, pages 94–101, Geneva, Switzerland.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Gerard de Melo and Stefan Siersdorfer. 2007. Multilingual text classification using ontologies. In Gianni Amati, editor, *Proc. 29th European Conference on Information Retrieval (ECIR 2007)*, volume 4425 of *Lecture Notes in Computer Science*, Rome, Italy. Springer.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Kevin Knight. 1993. Building a large ontology for machine translation. In *Proc. Workshop Human Language Technology*, pages 185–190.

Claudia Kunze and Lothar Lemnitzer. 2002. GermaNet - representation, visualization, application. In *Proc. LREC 2002*, pages 1485–1491.

Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. 2007. A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 68(3):267–276.

Márton Miháltz and Gábor Prószéky. 2004. Results and evaluation of Hungarian Nominal WordNet v1.0. In *Proceedings of the Second Global WordNet Conference*, Brno, Czech Republic. Masaryk University.

Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping WordNet to the suggested upper merged ontology. In *Proc. 2003 Intl. Conf. Information and Knowledge Engineering, Las Vegas, NV, USA*.

Akitoshi Okumura and Eduard Hovy. 1994. Building Japanese-English dictionary based on ontology for machine translation. In *Proc. Workshop on Human Language Technology*, pages 141–146.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proc. 1st Intl. Global WordNet Conference, Mysore, India*, pages 293–302.

John C. Platt, 1999. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA.

Frank Richter, 2007. *Ding Version 1.5*. `http://www-user.tu-chemnitz.de/~fri/ding/`.

German Rigau and Eneko Agirre. 1995. Disambiguating bilingual nominal entries against WordNet. In *Proc. Workshop 'The Computational Lexicon' at European Summer School Logic, Language & Information*.

Patanakul Sathapornrungkij and Charnyote Pluempitiwiriyawej. 2005. Construction of Thai WordNet lexical database from machine readable dictionaries. In *Proc. 10th Machine Translation Summit, Phuket, Thailand*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th Intl. World Wide Web Conference (WWW 2007)*. ACM Press.

Vladimir N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Piek Vossen, editor. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Springer.

Towards a Simple and Full-Featured Treebank Query Language

Jiří Mírovský
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
mirovsky@ufal.mff.cuni.cz

Abstract

Netgraph query language is a query system for linguistically annotated treebanks that aims to be sufficiently powerful for linguistic needs and yet simple enough for not requiring any programming or mathematical skill from its users. We provide an introduction to the system along with a set of examples how to search for some frequent linguistic phenomena. We also offer a comparison to the querying power of TGrep – a traditional and well known treebank query system.

1 Introduction

Searching in a linguistically annotated treebank requires a sophisticated tool, the more so the more complex the annotation is. Many users require (quite understandably) a simple and easy-to-learn tool, and yet they expect it to be satisfactorily powerful. It is obvious that there is a trade-off between simplicity of a query language and its searching power.

Netgraph has been designed to perform the searching with maximum comfort and minimum requirements on its users. Although it has been developed primarily for the Prague Dependency Treebank 2.0 (Hajič et al. 2006), it can be used with other treebanks too, both dependency and constituent-structure types.

In this paper, we present Netgraph query language and show how it can be used to search for some frequent linguistic phenomena. Afterwards, we try to compare the searching power of Netgraph query system to the power of traditional TGrep (Pito 1994), in order to check if it is at least the same. Thus, we set a lower boundary to the power of Netgraph query language. Therefore we concentrate on showing that TGrep does not outperform Netgraph and only mention what TGrep's flaws are, also because we know that there exist TGrep2, TigerSearch and other more recent tools. But we consider the power of TGrep the first step on the way of Netgraph towards "a full-featured searching tool". We plan to offer a comparison with the more recent tools in some future paper.

In *section 1* (after this introduction) we very briefly describe the Prague Dependency Treebank 2.0, just to make the examples in the subsequent text more understandable. Anyone familiar with this treebank may safely skip this subsection. In the next subsection we also mention in a few words the history of Netgraph and its properties as a tool.

In *section 2* we offer an introduction to the query language of Netgraph along with the idea of meta-attributes and what they are good for, and present several linguistically motivated examples of queries in the Prague Dependency Treebank. We also list all available meta-attributes.

In *section 3* we compare Netgraph query language to TGrep by translating TGrep predicates to Netgraph.

Finally, in *section 4* we offer some concluding remarks.

1.1 Prague Dependency Treebank 2.0

The Prague Dependency Treebank 2.0 (PDT 2.0, see Hajič et al. 2006, Hajič 2004) is a manually annotated corpus of Czech. It is a sequel to the

Prague Dependency Treebank 1.0 (PDT 1.0, see Hajič et al. 2001a, Hajič et al. 2001b).

The texts in PDT 2.0 are annotated on three layers - the morphological layer, the analytical layer and the tectogrammatical layer. The corpus size is almost 2 million tokens (115 thousand sentences), although "only" 0.8 million tokens (49 thousand sentences) are annotated on all three layers. By 'tokens' we mean word forms, including numbers and punctuation marks.

On the morphological layer (Hana et al. 2005), each token of every sentence is annotated with a lemma (attribute `m/lemma`), keeping the base form of the token, and a tag (attribute `m/tag`), keeping its morphological information. Sentence boundaries are annotated here, too.

The analytical layer roughly corresponds to the surface syntax of the sentence; the annotation is a single-rooted dependency tree with labeled nodes (Hajič et al. 1997, Hajič 1998). The nodes on the analytical layer (except for technical roots of the trees) also correspond 1:1 to the tokens of the sentences. The order of the nodes from left to right corresponds exactly to the surface order of tokens in the sentence. Non-projective constructions (that are quite frequent both in Czech (Hajičová et al. 2004) and in some other languages (see Havelka 2007)) are allowed. Analytical functions are kept at nodes (attribute `a/afun`), but in fact they are names of the dependency relations between a depending node (son) and its governing node (father).

The tectogrammatical layer captures the linguistic meaning of the sentence in its context. Again, the annotation is a dependency tree with labeled nodes. The correspondence of the nodes to the lower layers is more complex here. It is often not 1:1, it can be both 1:N and N:1. It was shown in detail in Mírovský (2006) how Netgraph deals with this issue.

Many nodes found on the analytical layer disappear on the tectogrammatical layer (such as functional words, prepositions, subordinating conjunctions, etc.). The information carried by these nodes is stored in attributes of the remaining (autosemantic) nodes and can be reconstructed. On the other hand, some nodes representing for example obligatory positions of verb frames, deleted on the surface, are regenerated on this layer.

The tectogrammatical layer goes beyond the surface structure of the sentence, replacing notions such as Subject and Object by notions like Actor, Patient, Addressee etc (see Hajičová 1998).

Attribute `functor` describes the dependency between a depending node and its governor and again is stored at the son-nodes. A tectogrammatical lemma (attribute `t_lemma`) is assigned to every node. Grammatemes are rendered as a set of 16 attributes grouped by the "prefix" `gram` (e.g. `gram/verbmod` for verbal modality).

The total of 39 attributes are assigned to every non-root node of the tectogrammatical tree, although (based on the node type) only a certain subset of the attributes is necessarily filled in.

Topic and focus (Hajičová et al. 1998) are marked (attribute `tfa`), together with so-called deep word order reflected by the order of nodes in the annotation (attribute `deepord`). It is in general different from the surface word order, and all the resulting trees are projective by the definition of deep word order.

To be complete (as much as possible in this short description), let us add that coreference relations between nodes of certain category types are captured (Kučová et al. 2003), distinguishing also the type of the relation (textual or grammatical). Each node has an identifier (attribute `id`) that is unique throughout the whole corpus. Attributes `coref_text.rf` and `coref_gram.rf` contain `id`s of coreferential nodes of the respective types.

1.2 Netgraph as a Tool

The development of Netgraph started in 1998 as a topic of Roman Ondruška's Master's Thesis (Ondruška 1998), and has been proceeding along with the ongoing annotations of the Prague Dependency Treebank 1.0 and later the Prague Dependency Treebank 2.0. Now it is a fully functional tool for complex searching in PDT 2.0.

Netgraph is a client-server application that allows multiple users to search the treebank on-line and simultaneously through the Internet. The server (written in C) searches the treebank, which is located at the same computer or local network. The client (written in Java2) serves as a very comfortable graphical user interface and can be located at any node in the Internet. It sends user queries to the server and receives results from it. Both the server and the client also can, of course, reside at the same computer. Authentication by the means of

login names and passwords is provided. Users can have various access permissions.

A detailed description of the inner architecture of Netgraph and of the communication between the server and the client was given in Mírovský, Ondruška and Průša (2002).

2 Netgraph Query Language

In this section we give an introduction to the Netgraph query language. We show on a series of examples how some frequent linguistic phenomena can be searched for.

2.1 The Query Is a Tree

The query in Netgraph is a tree that forms a subtree in the result trees. The treebank is searched tree by tree and whenever the query is found as a subtree of a tree (we say the query and the tree match), the tree becomes part of the result. The result is displayed tree by tree on demand. The query can also consist of several trees joined either by `AND` or `OR` relation. In that case, all the query trees at the same time (or at least one of the query trees, respectively) are required to match the result tree.

The query has both a textual form and a graphical form. For lack of space, we will use its textual form in this paper. However, each textual query has its full graphical counterpart, which is always much more transparent.

The syntax of the language is very simple. In the textual form, square brackets enclose a node, attributes (pairs `name=value`) are separated by a comma, quotation marks enclose a regular expression in a value. Parentheses enclose a subtree of a node, brothers are separated by a comma. In multiple-tree queries, each tree is on a new line and the first line contains only a single `AND` or `OR`. Alternative values of an attribute, as well as alternative nodes, are separated by a vertical bar. It almost completes the description of the syntax, only one thing (references) will be added in the following subsection.

The simplest possible query (and probably of little interest on itself) is a simple node without any evaluation: `[]`. It matches all nodes of all trees in the treebank, each tree as many times as how many nodes there are in the tree. Nevertheless, we may add conditions on its attributes, optionally using regular expressions in values of the attributes. Thus we may search e.g. for all nodes that are Subjects and nouns but not in first case:

```
[afun=Sb, m/tag="N...[^1].*"].
```

We may notice here that regular expressions allow the first (very basic) type of negation in queries.

More interesting queries usually consist of several nodes, forming a tree structure. The following example query searches for trees containing a Predicate that directly governs a Subject and an Object:

```
[afun=Pred]([afun=Sb],[afun=Obj]).
```

Please note that there is no condition in the query on the order of the Subject and the Object, nor on their left-right position to their father. It does not prevent other nodes to be directly governed by the Predicate either.

2.2 Meta-Attributes

This simple query language, described briefly in only a few examples, is quite useful but not powerful enough. There is no possibility to set a real negation, no way of restricting the position of the query in the result tree or the size of the result tree, nor the order of nodes can be controlled. To allow these and other things, meta-attributes have been added to the query system.

Meta-attributes are not present in the corpus but they pretend to be ordinary attributes and the user uses them the same way like normal attributes. Their names start with an underscore. There are eleven meta-attributes, each adding some power to the query language, enhancing its semantics, while keeping the syntax of the language on the same simple level. We present several of the meta-attributes in this subsection, some others will be presented in the subsequent section, when they are needed. A list of all meta-attributes is presented in the next subsection.

Coordination is a frequent phenomenon in languages. In PDT (and in most other treebanks, too) it is represented by a coordinating node. To be able to skip (and effectively ignore) the coordination in the queries, we have introduced the meta-attribute `_optional` that marks an optional node. The node then may but does not have to be in the result. If we are interested, for example, in Predicates governing Objects, we can get both cases (with coordination and without it) in one query using this meta-attribute:

```
[afun=Pred]([afun=Coord,_option-
al=1]([afun=Obj])).
```

The Coordination becomes optional. If there is a node between the Predicate and its Object in the result tree, it has to be the Coordination. But the Object may also be a direct son of the Predicate, omitting the optional Coordination.

There is a group of meta-attributes of rather technical nature. They allow setting a position of the query in the result tree, restricting the size of the result tree or its part, and restricting number of direct sons of a node. Meta attribute `_depth` controls the distance of a node from the root (useful when searching for a phenomenon in subordinated clauses, for example), `_#descendants` controls number of nodes in the subtree of a node (useful e.g. when searching for „nice“ small examples of something), `_#sons` controls number of (direct) sons of a node.

Controlling number of direct sons (mainly in its negative sense) is important for studying valency of words (Hajičová and Panevová 1984). The following example searches on the tectogrammatical layer of PDT. We want a Predicate that governs directly an Actor and a Patient and nothing else (directly):

```
[functor=PRED,_#sons=2]([func-
tor=ACT],[functor=PAT]).
```

If we replaced PAT with ADDR, we might search for errors in the evaluation, since the theory forbids Actor and Addressee being the only parts of a valency frame.

So far, we could only restrict number of nodes. But we often want to restrict a presence of a certain type of node. We want to specify that there is not a node of a certain quality. For example, we might want to search (again on the tectogrammatical layer) for an Effect without an Origo in a valency frame. The meta-attribute that allows this real type of negation is called `_#occurrences`. It controls the exact number of occurrences of a certain type of node, in our example of Origos:

```
[functor=PRED]([functor=EFF],[fu
nctor=ORIG, _#occurrences=0]).
```

It says that the Predicate has at least one son – an Effect, and that the Predicate does not have an Origo son.

There is still one important thing that we cannot achieve with the meta-attributes presented so far. We cannot set any relation (other than dependency) between nodes in the result trees (such as order, agreement in case, coreference). All this can be done using the meta-attribute `_name` and a system of references. The meta-attribute `_name` simply names a node for a later reference from other nodes.

Curly brackets enclose a reference to a value of an attribute of the other node (with a given name) in the result tree. This, along with the dot-referencing inside the reference and some arithmetic possibilities, completes our description of the syntax of the query language from subsection 2.1.

In the following example (back on the analytical layer and knowing that attribute `ord` keeps the order of the node (~ token) in the tree (~ sentence)), we search for a Subject that is on the right side from an Object:

```
[afun=Pred]([afun=Sb,ord>{N1.ord
}],[afun=Obj,_name=N1]).
```

We have named the Object node `N1` and specified that ord of the Subject node should be bigger than `ord` of the `N1` node. If we used `ord>{N1.ord}+5`, we would require them to be at least five words apart.

2.3 List of All Meta-Attributes

To complete our description of Netgraph query language, we present all available meta-attributes in one list, along with a short description:

`_transitive`

This meta-attribute defines a transitive edge. It has two possible values: `true` means that a node may appear anywhere in the subtree of its query-father, `exclusive` means, in addition, that the transitive edge cannot share nodes in the result tree with other exclusively transitive edges.

`_optional`

It defines an optional node. It may but does not have to appear in the result. However, if there is a node in the result at this particular place (father in grandfather-father-son hierarchy), it must be the one defined in the query. Depending on the value of this meta-attribute, one or more nodes may be skipped. A special value `true` skips an unlimited chain of the specified nodes.

`_#sons`

It defines an exact number of sons of a query-node in the result tree.

`_#hsons`

It defines an exact number of hidden sons of a query-node in the result tree. Hidden nodes are especially marked nodes in the tree that provide connection to the information on the lower layers of annotation. They are useful when the relation between nodes at different layers is not 1:1. A detailed description of the system of hidden nodes was given in Mírovský (2006).

`_#descendants`

This meta-attribute defines an exact number of all descendants of a node (number of nodes in its subtree), excluding the node itself.

`_#lbrothers`

This meta-attribute defines an exact number of left brothers of a node.

`_#rbrothers`

Similarly, it defines an exact number of right brothers of a node.

`_depth`

It defines a distance between a node and a root in the result tree.

`_#occurrences`

This meta-attribute specifies an exact number of occurrences of a particular node at a particular place in the result tree. It controls how many nodes of the kind can occur in the result tree as sons of the father of the node (including the node itself). It can be combined with meta-attribute `_transitive` for transitive meaning of the above definition.

`_name`

It names a node for references to values of its attributes in the result tree.

`_sentence`

The value of this meta attribute is simply the sentence the result tree belongs to in its linear form. It can be used for linear searching in the sentence (using regular expressions).

3 Comparison to TGrep

In this section, we compare the query language of Netgraph to the query language of TGrep, in order to show that the power of Netgraph query language is at least the same as the power of TGrep. We also show at the end that Netgraph has a greater power.

In subsection 3.1 we compare the ability of expressing an evaluation of a node. In the next two subsections (3.2 and 3.3) we translate TGrep positive and negative predicates to Netgraph expressions. In subsection 3.4 we give an example of Netgraph expressions that cannot be searched for in TGrep.

3.1 Node Evaluation

TGrep is a one-attribute searcher. Each node is supposedly labeled only either by a non-terminal symbol or a token. Netgraph, on the other hand, can deal with multiple attributes and set conditions on them separately and even form groups of them that are labeled differently (so called "alternative nodes"). Leaving this aside, we can say that Netgraph has (at least) the same expressing power in the sense of node values as TGrep does, as both tools allow using regular expressions and set alternative values. Thus, we can almost simply repeat the example of a search pattern from TGrep manual:

in TGrep:

```
/^[Cc]hild.*$/|kid|youngster
```

in Netgraph:

```
"[Cc]hild.*"|kid|youngster
```

Netgraph regular expressions are automatically anchored and are enclosed in quotation marks. The complete query in Netgraph in the text form would then be (it also has to be "escaped" in the text form, though not in the graphical form):

```
[token="\[Cc\]hild.*"|kid|young-
ster]
```

The wildcard represented by two underscores in TGrep is reproducible in Netgraph by not specifying any attribute at the node: `[]`.

3.2 Tree Structure

The close similarity between Netgraph and TGrep in expressing node evaluations disappears completely when it comes to defining relations between nodes. Here, these two tools have quite a different approach. The main difference is that TGrep uses predicates to express dependency between nodes, while Netgraph expresses dependency directly in the syntax of the query. In this subsection, we try to match TGrep positive predicates with similar constructions in Netgraph. We take predicates (relationships between nodes) from

TGrep manual one by one and translate them to equivalent Netgraph expressions.

The first line of each example (starting with T) always shows the expression in TGrep, while the second line (starting with N and occasionally followed by other lines) shows the equivalent expression in Netgraph.

A immediately dominates B:

```
T: A < B
N: [A]([B])
```

B is the X-th son of A:

```
T: A <X B
N: [A]([B,_#lbrothers=X-1])
```

We use meta-attribute _#lbrothers here, which specifies how many left brothers a node has. X-th to last son is similar, we only use meta-attribute _#rbrothers (number of right brothers).

A dominates B (A is dominated by B similarly):

```
T: A << B
N: [A]([B,_transitive=true])
```

Meta-attribute _transitive defines the father edge as transitive.

B is the leftmost (rightmost) descendant of A:

```
T: A <<, B
N:
[A]([B,_transitive=true,_name=N1],
[_transitive=true,ord<{N1.ord},
_#occurrences=0]).
```

B is a transitive descendant of A and there is no transitive descendant of A that has smaller ord than B. Rightmost descendant is similar (ord<{N1.ord}).

A immediately precedes B:

```
T: A . B
N: AND
[A,_name=N1]
[B,ord={N1.ord+1}]
```

Since we generally do not know what dependency relation between the two nodes is, we must define them as two separate trees in a multiple-tree query (another possibility is to use a wildcard and two transitive sons). A precedes B is similar, we only use a different expression in the second tree:

```
N: [B,ord>{N1.ord}]
```

A and B are brothers:

```
T: A $ B
N: [][A],[B])
```

We use the wild card here since we generally do not know anything about the father (we only know that there must be one).

A and B are brothers and A immediately precedes B:

```
T: A $. B
N: [][A,_name=N1][B,
_#brothers={N1._#brothers}+1])
```

We have to use meta-attribute _#brothers here instead of attribute ord, because there may be other nodes (not brothers of A and B) in between them in left-right order of nodes. On the other hand, if we wanted to take the other nodes into account, we might use attribute ord.

Of course, things get more complex when we start combining these expressions. We believe that in Netgraph the complex expressions remain well readable. Sometimes we may be lucky and have a convenient meta-attribute at our disposal, just like in the following example, taken again from TGrep manual, which specifies all nodes A that dominate either two or three sons:

```
T: A <2 __ !<4 __
N: [A,_#sons=2|3]
```

3.3 Negation

Netgraph's way of specifying relations between nodes, especially their dependency, is primarily positive and it has some difficulty expressing negative relations. For this reason, it is sometimes not easy or even possible to match directly and exactly TGrep negative expressions without "saying" something positive about the nodes, too.

A does not immediately dominate B:

```
T: A !< B
N: [A]([B,_#occurrences=0]).
```

B is not the X-th son of A:

```
T: A !<X B
N: A([B,_#lbrothers!=X-1])
```

But note that it also means that B is a son of A. Using meta-attribute _#occurrences again, we may have another try on this example with a different meaning:

```
N: [A]([B,_#lbrothers=X-1,_#occur
rences=0])
```

Here, B still may be a son of A, but not necessarily, and in any case not the X-th one.

A does not dominate B:

```
T: A !<< B
N: [A]([B,_transitive=true,_#occur
rences=0])
```

B is not the leftmost descendant of A:

```
T: A !<<,B
```

This again must be considered in two separate cases: positive and negative. If we only want to say that the leftmost descendant of A has another property than B, the query in Netgraph would be:

```
N: [A]([!B,_transitive=true,
_name=N1],[_transitive=true,
ord<{N1.ord},_#occurrences=0]).
```

On the other hand, if we want to say that B is a descendant of A that is not the leftmost one, the query would be:

```
N: [A]([B,_transitive=true,
_name=N1],[ord<{N1.ord},_#occur-
rences>=1,_transitive=true])
```

A does not immediately precede B:

```
T: A !. B
N: AND
[A,_name=N1]
[!B,ord={N1.ord+1}]
```

Which is very similar to the positive case from the previous subsection. Note that it also means that there is a directly subsequent node !B in the result tree (a node that does not have B-property).

A does not precede B:

```
T: A !.. B
```

Just like before, two possible interpretations of this expression lead to two different realizations in Netgraph. The positive meaning is quite simple – A does not precede B is equal to B precedes A (since nodes cannot have the same left-right order). The negative meaning (there is A that is not followed by B) would be translated:

```
AND
[A,_name=N1]
[B,ord>{N1.ord},_#occurrences=0]
```

A is not a brother of B:

```
T: A !$ B
N: [][A],[B,_#occurrences=0])
```

If we also wanted to use B positively in the query, we might add another tree of a multiple-tree query.

It is not true that A $. B (similarly A !$.. B)

```
T: A !$. B
```

Many possible interpretations of this expression lead to many different realizations of the equivalent Netgraph query. We will not show all of them (they are all similar to the previous queries) but only choose the most direct one, B is a brother of A but does not immediately follow A:

```
N: [][A,_name=N1],
[B,_#lbrothers!={N1._#lbrothers}+1
])
```

3.4 The Other Way

Since TGrep always searches for one pattern only, it cannot reproduce multiple-tree queries from Netgraph, combined with expression OR. Meta-attribute `_optional` also represents a type of OR-expression on the tree structure and even the simple example given in subsection 2.2 cannot be reproduced in TGrep:

```
[afun=Pred]([afun=Coord,_option-
al=1]([afun=Obj])).
```

4 Conclusion

We have presented Netgraph query language on a set of linguistically motivated examples. We have compared Netgraph query power to the power of TGrep query language in order to show that it is not lesser, by translating all TGrep predicates to expressions in Netgraph. We have also shown that some Netgraph expressions cannot be translated to TGrep.

Many constructions in Netgraph seem more complicated than respective expressions in TGrep. The reason is that we matched TGrep predicates. It is clear that any other system that uses a different set of predicates cannot be as straightforward as TGrep in mimicking these predicates. It is sufficient for our purpose that the translation is possible.

We can conclude that Netgraph query language is at least as strong as TGrep query language. The impossibility of translating OR-expressions from Netgraph to TGrep shows that Netgraph query language is stronger than TGrep query language.

Acknowledgment

The research reported in this paper was supported by the Grant Agency of the Academy of Sciences of the Czech Republic, project IS-REST (No. 1ET101120413).

References

Hajič J. et al. 2006. Prague Dependency Treebank 2.0. *CD-ROM LDC2006T01, LDC, Philadelphia, 2006*.

Pito R. 1994. TGrep Manual Page. *Available from http://www.ldc.upenn.edu/ldc/online/treebank/*

Hajič J. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. *Jazykovedný ústav Ľ. Štúra, SAV, Bratislava, 2004*.

Hajič J., Vidová-Hladká B., Panevová J., Hajičová E., Sgall P., Pajas P. 2001a. Prague Dependency Treebank 1.0 (Final Production Label). *CD-ROM LDC2001T10, LDC, Philadelphia, 2001.*

Hajič J., Pajas P. and Vidová-Hladká B. 2001b. The Prague Dependency Treebank: Annotation Structure and Support. *In IRCS Workshop on Linguistic databases, 2001, pp. 105-114.*

Hana J., Zeman D., Hajič J., Hanová H., Hladká B., Jeřábek E. 2005. Manual for Morphological Annotation, Revision for PDT 2.0. *ÚFAL Technical Report TR-2005-27, Charles University in Prague, 2005.*

Hajič J. et al. 1997. A Manual for Analytic Layer Tagging of the Prague Dependency Treebank. *ÚFAL Technical Report TR-1997-03, Charles University in Prague, 1997.*

Hajič J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. *In Issues of Valency and Meaning, Karolinum, Praha 1998, pp. 106-132.*

Hajičová E., Havelka J., Sgall P., Veselá K., Zeman D. 2004. Issues of Projectivity in the Prague Dependency Treebank. *MFF UK, Prague, 81, 2004.*

Havelka J. 2007. Beyond Projectivity: Multilingual Evaluation of Constraints and Measures on Non-Projective Structures. *In Proceedings of ACL 2007, Prague, pp. 608-615.*

Hajičová E, Panevová J. 1984. Valency (case) frames. *In P. Sgall (ed.): Contributions to Functional Syntax, Semantics and Language Comprehension, Prague, Academia, 1984, pp. 147-188.*

Mírovský J. 2006. Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0. *In Proceedings of TLT 2006, Prague, pp. 211-222.*

Hajičová E. 1998. Prague Dependency Treebank: From analytic to tectogrammatical annotations. *In: Proceedings of 2nd TST, Brno, Springer-Verlag Berlin Heidelberg New York, 1998, pp. 45-50.*

Hajičová E., Partee B., Sgall P. 1998. Topic-Focus Articulation, Tripartite Structures and Semantic Content. *Dordrecht, Amsterdam, Kluwer Academic Publishers, 1998.*

Kučová L., Kolářová-Řezníčková V., Žabokrtský Z., Pajas P., Čulo O. 2003. Anotování koreference v Pražském závislostním korpusu. *ÚFAL Technical Report TR-2003-19, Charles University in Prague, 2003.*

Ondruška R. 1998. Tools for Searching in Syntactically Annotated Corpora. *Master Thesis, Charles University in Prague, 1998.*

Mírovský J., Ondruška R., Průša D. 2002. Searching through Prague Dependency Treebank - Conception and Architecture. *In Proceedings of The First Workshop on Treebanks and Linguistic Theories, Sozopol, 2002, pp. 114--122.*

Minimally supervised lemmatization scheme induction through bilingual parallel corpora

Taesun Moon and Katrin Erk
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
tsmoon,katrin.erk@mail.utexas.edu

Abstract

We present a lemma induction scheme on a target language through minimally supervised alignment and transfer methods utilizing English-to-German parallel corpora. Compared to previous alignment and transfer approaches, the approach outlined here increases computational efficiency and significantly reduces the level of supervision necessary in inducing clusters of inflectional forms. Furthermore, we increase our search field to include not only verbs but also nouns and adjectives in the target language, and achieve comparable results to previous unsupervised monolingual methods.

1 Introduction

Cross-language projection of linguistic information through alignment and transfer methods using parallel corpora has been used for a variety of tasks and purposes such as deriving the syntactic structure of a target language (Wu, 1997), extracting paraphrases (Pang et al., 2003; Bannard and Callison-Burch, 2005), extracting bilingual knowledge(Shin et al., 1996), or semantic disambiguation (Diab, 2000). Among these, one group of approaches has focused on inducing basic NLP tools such as POS taggers, noun chunkers, and morphology analyzers for a given target language (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001; Yarowsky and Ngai, 2001; Drábek and Yarowsky, 2005; Ozdowska, 2006; Moon and Baldridge, 2007). The latter approaches take note of the fact that many of the languages of the world are underdocumented and resource challenged, and that there is a need to provide rudimentary but robust tools to assist in the process of documentation and analysis.

The study outlined here is in line with this basic premise. Using sentence-aligned parallel texts from the Europarl Corpus (Koehn, 2005), with English as the source and German as the target, we induce a lemmatization scheme over the target language through alignment and transfer methods. The same problem has been dealt with only once before in Yarowsky et al.(2001), which treated verbs in Czech and French. We instead propose a different approach which foregoes some of their basic assumptions as well as a probabilitistic model based on those assumptions and instead focuses on methods for reducing the search space of candidate lemmata, and expands the lemmatization to incorporate not just verbs but also nouns and adjectives. With an overall token precision of 0.836, we achieve results comparable to other unsupervised methods. Given that the induction of lemmatization schemes is an important stepping stone in building other fundamental NLP tools such as lemmatizers, POS taggers, and parsers, we believe the aims and results of this study can provide useful insights as well as critical data in this process of accumulating tool sets.

In this paper, after a review of related work including a discussion of the models and assumptions in Yarowsky et al., and a presentation of the data sets and programs we will be using, we outline and present our own approach which, despite failing to best the results of Yarowsky et al., show that robust results are possible in spite of a significantly

reduced level of supervision, even when the target word categories are expanded to include not only verbs but also nouns and adjectives. In section 5, we provide the results of our attempt to reimplement Yarowsky et al., present our own results, and evaluate them according to two criteria, one of which is novel in its assessment of hard clustering tasks such as the lemmatization task attempted here and is more methodologically sound given the nature of the task.

2 Related Work

Unsupervised monolingual morphology segmentation is a topic that has been tackled many times in the literature (Goldsmith, 2001; Sassano, 2001; Goldwater, 2006; Hammarström, 2006; Creutz and Lagus, 2007). Though such approaches generally manage to provide relatively reliable segmentation schemes with precisions between the ranges of 0.8 and 0.9, it is difficult to generalize beyond the segmentation of individual word types to how they relate to the POS categories in a given language or its syntax. We show in this paper that alignment and transfer methods based on utilizing the linguistic metadata of a well-documented language for the analysis of another can provide concrete motivation for limiting the search space for potential clusterings of inflected forms and can also impose higher-level syntactic constraints. This will result in an analysis that is less dependent on the quirks of orthographic similarities.

Yarowsky et al.(2001) introduced the method of lemmatization scheme induction through alignment and transfer methods. It forms part of a larger group of studies that focus on the use of bilingual corpora to induce NLP tools for a target language (Shin et al., 1996; Wu, 1997; Diab, 2000; Yarowsky et al., 2001; Drábek and Yarowsky, 2005; Ozdowska, 2006). In this study, a core algorithm in the induction of lemmatization schemes in a target language is the transitivity function, an approach based on the intuition that if one lexeme and another lexeme in the target language have been aligned with more lemmas in the source language than with some other lemma, the more likely it is that the two words can be grouped together under some meaningful cluster. They use the following probabilistic model:

$$P(T_{lemma}|T_{infl}) = \sum_i P(T_{lemma}|S_{lemma_i})P(S_{lemma_i}|T_{infl}) \quad (1)$$

In this approach, the probability that a target lemma T_{lemma} will be the lemma of an inflected token in the target T_{infl} is estimated by summing over the probability of T_{lemma} given a lemma S_{lemma_i} in the source multiplied by the probability of the source lemma given T_{infl} for all the lemmas in the source. The transitive links used here will increase the likelihood of $P(T_{lemma}|T_{infl})$ the more often they co-occur over all source lemmas which provide a link between the two.

The major limitation of this approach is that it requires a pre-selected list of lemmata in the target language. Though it is possible to modify the model and implementation so that no assumptions are necessary regarding which word types in the target are lemmata, or "dictionary entry forms", and which are inflected forms, such a modification comes at the cost of considerable time complexity. Needless to say, manually selecting a set of target lemmata as has been done in this study is a step which significantly increases the level of supervision.

A second limitation of this approach (discussed with examples in Section 5) is that the transitivity function when implemented without any assumptions regarding lemmata casts a very wide net, favoring retrieval over precision. Even when implemented on a manually selected set of target lemmata, Yarowsky et al. impose a empirically determined threshold (which is unspecified) on the transitivity function to limit the size of the sets of candidate inflectional forms which have been associated with some candidate target lemma. It is not discussed whether this same threshold is applicable across a wide specturm of languages, and further investigation might reveal that a case-by-case inspection of the data is required in each instance to determine this threshold.

With this approach, they post a precision of 0.992 and a recall of 0.994 over word tokens for the 12M word French Hansards using the alignment method alone. However, it should be noted that the induction was performed for only verbs in the target language and that the study had implemented a POS tagger induced through similar minimally supervised means.

They gain a small increase in precision and a substantial increase in retrieval over target word types by augmenting the above approach with a trie based search and a backoff model based on Levenshtein distance and distributional similarity measure. With the aid of these methods, they increase the general level of precision over all their target corpora to an almost insuperable 0.99 and a retrieval of 1.00. The latter augmentative approaches are outlined in more detail in Yarowsky and Wicentowsky (2000).

3 Data and Alignment

3.1 Europarl Parallel Corpus

The German and English sections of the Europarl parallel corpus (Koehn, 2005) were used in this study. The Europarl parallel corpus is a collection of texts in 11 languages extracted from the proceedings of the European parliament with each text comprising some 25 to 30 million words. Any two texts from this corpus are mutually parallel.

To enhance the accuracy of the parsing and alignment tasks, the parallel corpus was further trimmed to English sentences of less than 45 words in length. This reduced the size of the English and German corpus to roughly 17 million words each.

3.2 Lemmatization and POS tagging of source text

POS tagging for the English text was done with the maximum entropy based C&C tagger (Curran and Clark, 2003), which was trained on the Wall Street Journal of the Penn Treebank. The POS tagged source text was then supplied to the lemmatizer, Morpha (Minnen et al., 2001), a finite state morphology analyzer, whose only requirement for prior POS tagged data is that verbal tags are headed by a V and noun tags other than proper nouns are headed by an N. Such knowledge of the word category of a lexeme is necessary in enhancing the performance of Morpha.

3.3 Word Alignment

Word alignment between the two texts was achieved with GIZA++ (Och and Ney, 2003). The alignment was made with English as the source and German as the target. In this stage, the parallel corpus was further reduced to three-quarters of the trimmed corpus derived in the stage outlined above. It was a process recommended in Yarowsky et al. (2001) to reduce undue noise in the alignment model, and so alignments with a confidence measure in the lower 25% of the parallel corpus were removed from consideration for this study.

3.4 TIGER Treebank Corpus

The TIGER Treebank (Brants and Hansen, 2002) corpus was used as the evaluation corpus on which to test lemmatization schemes. The corpus, which is currently at version 2.1, is a collection of German newspaper text gathered from the Frankfurter Rundschau and consists of app. 900,000 tokens. It is annotated with POS tags and lemmata for terminal nodes and has been manually annotated for syntactic information. The use of this corpus also allowed us to evaluate how well the scheme induced from one domain would translate to another.

4 Approach

Our approach wholly does away with the transitivity function by aggressively culling the search space for candidate lemmata and candidate inflected forms. First, we limit the set of candidate lemmata to the word types in the target language which have the greatest possibility of being associated with some lemma in the source language. With this candidate lemma, we generate one set of lemmata to inflected form mappings by limiting the linkages to those source lemmata and target word type associations which exceed a manually determined probability threshold. We generate a second set of mappings from a candidate lemma to a set of target word types which has been limited to those which have been observed in alignment with a source lemma and then further reduced through an automatically induced edit distance threshold.

4.1 Lemmatization candidate trimming

Using the word based alignment output from GIZA++, we obtained the conditional likelihood estimates from the target text:

$$P(\ell_s T_s | w_t) \quad (2)$$

$$P(w_t | \ell_s T_s) \quad (3)$$

where subscripts s and t are source and target texts, respectively, ℓ and T are lemma and POS tag, respectively, and w is a word type in the target language. ℓ_s is an element of the set Λ_s which is the set of all lemmata observed in the source language and w_t is an element of the set W_t which is the set of all types observed in the target language. In comparison to the two previous attempts to lemmatize a target language through alignment and transfer methods (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001), we expand the set of POS tags from verbs to incorporate adjectives and nouns as well, in short all the content word categories in English except for the adverbs.

Therefore, all lemmata in the English source had to be considered with their respective POS tags, considering that many lemmata in English can be ambiguous with regard to word category when judged on their surface form alone. Hereinafter, to simplify notation, all source lemma arguments in functions shall be assumed to also be tagged with relevant POS information. As such, the above equations are equivalent to

$$P(\ell_s|w_t) \tag{4}$$

$$P(w_t|\ell_s) \tag{5}$$

Also, note that words in the aligned target text are merely assumed to be a word type in the most general sense, since no assumptions can be made at this point whether a particular word form observed in the target language is the inflected form of some lemma or is itself the general "dictionary entry form".

In the estimation of the probablities in (4) and (5), we make an unjustified but practical decision to limit the set of target word types under examination to those which have string lengths of four or longer. This was mainly due to the fact that the Levenshtein edit distance algorithm is incapable of calculating meaningful scores when the strings being compared are both very short.

To limit the search space, we build two mapping tables, one from the target word types to the source lemmata and another from the source lemmata to the target word types.

The mapping from the target to the source, $TS : W_t \rightarrow \Lambda_s$, is built by

$$TS(w_t) = \ell_s \text{ iff } P(\ell_s|w_t) > 0.75 \tag{6}$$

The mapping from the source to the target, $ST : \Lambda_s \rightarrow W_t$ is built by

$$ST(\ell_s) = \arg\max_{w_t} P(w_t|\ell_s) \tag{7}$$

The mapping from target to source TS is ensured to be unambiguous since the probability threshold for assigning a mapping from w_t to ℓ_s is 0.75. This threshold value was selected after an initial examination of the data extracted from the Europarl corpus; and given its high threshold, it is expected to generate high confidence candidates regardless of the target language. However, future studies will need to examine methods of automating the threshold extraction procedure.

Using the two mappings TS and ST, we will automatically determine a minimal Levenshtein edit distance threshold by comparing the edit distance between all possible W_t to W_t mappings,

$$ST(TS(w_t)) = w'_t \tag{8}$$

where $w_t, w'_t \in W_t$. The mapping obtained here will be necessary for limiting the search space for the first set of candidate lemma to candidate inflectional form mappings.

Declare: $a[0 \ldots n]$
1: **for** j from 0 to n **do**
2: $\quad a[j] := 0$
3: **end for**
4: **for all** $w_t \in W_t$ **do**
5: $\quad$ **if** $TS(w_t) \neq$ NONE **then**
6: $\quad\quad w'_t := ST(TS(w_t))$
7: $\quad\quad d :=$ edit_distance(w_t,w'_t)
8: $\quad\quad$ **if** $d < n + 1$ **then**
9: $\quad\quad\quad a[d] := a[d] + 1$
10: $\quad\quad$ **end if**
11: $\quad$ **end if**
12: **end for**
13: **return** $min(a[0 \ldots n])$

Figure 1: Algorithm for computing edit distance threshold

The specific algorithm for computing the edit distance threshold is laid out in Figure 1. We obtain the edit distance for every w_t, w'_t pair in (8), and keep count of how many times each edit distance score

was observed (which is stored in an array a of length n in the algorithm; in this case, we used an array of length 9). Finally, the edit distance threshold is determined to be the minima among the frequency counts by edit distance score. The actual frequencies can be observed in Figure 2, the graph of which approximates a convex function. Furthermore, even if the number of edit distance scores we keep track of is increased to include all edit distance scores, it is evident that a score and its frequency count will continue to increase until reaching some asymptotic upper limit for all real-word data. Therefore, though the highest edit distance score we maintain a frequency count of is 9, there is no possibility that the frequency count will decrease at some point above that score. The intuition behind the approach is that two target words which have an edit distance beyond a certain threshold is more likely to be noise and those which do not exceed it will be related within some inflectional paradigm; and that this threshold exists at the minima of the frequency counts.

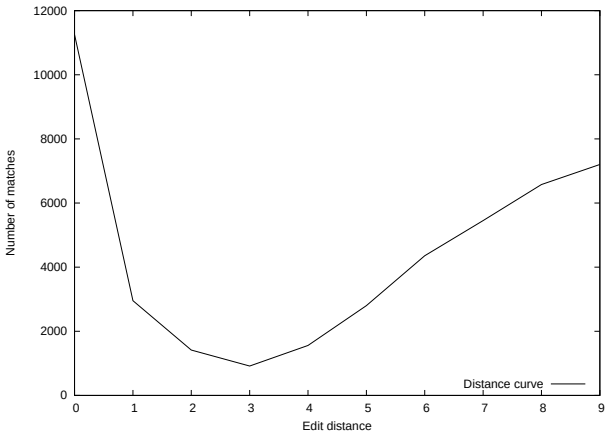


Figure 2: Extraction of Levenshtein edit distance threshold

4.2 Candidate set induction

We induce our first set of lemma group candidates as follows. First, we generate a mapping M from a source lemma ℓ_s to a set of target word types $\Omega_t \subset W_t$ where

$$\Omega_t = \{w_t | P(w_t|\ell_s) > 0\}$$

With this mapping

$$M(\ell_s) = \Omega_t$$

we further trim Ω_t by pegging the lemma candidate as $ST(\ell_s)$ (see equation (7)) and removing all the elements in Ω_t which have a Levenshtein edit distance score from $ST(\ell_s)$ greater than the distance threshold 3 (obtained through the algorithm in Figure 1), resulting in Ω'_t, a subset of Ω_t.

Thus, we have obtained a set of lemma candidates Λ_t in the target language

$$\Lambda_t = \{\ell_t | \forall \ell_s \in \Lambda_s, ST(\ell_s) = \ell_t\}$$

and a set of inflections associated with each ℓ_t in Λ_t

$$C_1(\ell_t) = \Omega'_t$$

Furthermore, the candidate lemma ℓ_t inherits the POS tag from the source language, so that ℓ_t is also specified for whether it is an adjective, noun, or verb.

A second candidate set, or a mapping from candidate lemma to candidate inflected forms, is induced by trimming the mapping TS to a subset of mappings where if the length of the common substring between the input and the output is less than 4, it is removed. However, the common substring in this case is not the longest common substring assumed in general, but merely the common substring from the beginning of each string being compared.

The justification for this is as follows. A very simple assumption can be made that a language will be either prefixal or suffixal in its inflectional system. By implementing two tries over the entire set of word types in the target language W_t, one trie starting from the beginning of the strings and another starting from the end of the strings[1], we can compare how many terminal nodes there are for the forward trie and the reverse trie, the intuition being that the more terminal nodes a particular trie has, the less likely it is that morphological affixation occurs at the terminal nodes of that trie. In the case of our study, it was found that the forward trie had 898 terminal nodes whereas the reverse trie had 4387 terminal nodes. Hence, we come to the simplified conclusion that the target language was suffixal rather than prefixal in generating inflected forms.

The second candidate lemma to candidate inflection mapping, unlike the first candidate mapping, is not from a word type to a set, but from a word type

[1]Again, the word types that were submitted to the trie were restricted to those whose length was greater than 3

to a word type. We define the second candidate mapping C_2 as follows:

1: **for all** $w_t \in W_t$ **do**
2: **if** $ST(TS(w_t)) \neq \texttt{NONE}$ **then**
3: $w'_t := ST(TS(w_t))$
4: **if** $CS(w_t, w'_t) < 3$ **then**
5: $C_2(w_t) = w'_t$
6: **end if**
7: **end if**
8: **end for**

where *CS* is a function on two strings which returns an integer value of the longest common substring starting from the beginning of the two arguments and $ST(TS(w_t))$ is the mapping stated in (8).

Finally, we combine the two candidate mappings into a final candidate mapping C which is a relation from a word type to a set of word types. If there are coinciding ℓ_t in C_1 and C_2, then the output of C_2 is merged into the set generated by C_1. Otherwise, candidates are simply added to the mapping C.

5 Results and Evaluation

5.1 An examination of the transitivity function

In our implementation of the transitivity function in (1), we modified the model so that it would not make any assumptions about which words in the target are lemmata and which are not. However, this revealed itself to be computationally too intense in terms of time complexity. When we limited the candidate set of target text lexemes to about 100 and the set of source text lemmata to 50000, it took 60 minutes to complete the computation. It would have been impossible to expand the set of target text lexemes to the 50000 word types that we had. When we reduced the number of source text lemmata to a manageable 1000 words, we were confronted with the problem of sparse data and the function was not able to properly link candidate lexemes. Finally, we tried the option of reducing the set of source lemmata and target lexemes to those which had been observed in transitive verb/direct object relations in the source, the syntactic relations of which were obtained through the C&C tagger (Curran and Clark, 2003).

A small subsample of the results can be observed in Figure 3. In addition to the examples observed in the subsample, the amount of noise in the results in general were excessive and ultimately unfit for inducing lemmatization schemes. In addition to manually defining a set of target lemmata, Yarowsky et al. (2001) used a manually set threshold for the transitivity values obtained through Equation 1 to remove the unfit pairings between candidate lemma and candidate inflected form. While such a threshold over this data might have reduced the level of noise, in the end, it would have been prohibitively time consuming to achieve an enhancement in retrieval or precision over our data set.

5.2 Revised approach

There were 193582 word types in the German portion of the Europarl corpus. From this set W_t, 15945 lemma candidates were induced after applying the culling outlined in section 4. These lemma candidates were mapped to a total of 29056 candidate inflected forms, an average of 1.8 inflectional candidates to a lemma candidate.

Evaluation was conducted using two separate measures. One was over the tokens observed in the TIGER corpus (Figure 4) and another was over types (Figure 5).

	ADJ	N	V	OVERALL
Precision	0.711	0.903	0.718	0.836
Recall	0.277	0.330	0.080	0.267
F-Score	0.399	0.483	0.144	0.405

Figure 4: Scores by tokens and POS tag

	ADJ	N	V	OVERALL
Precision	0.711	0.795	0.840	0.772
Recall	0.822	0.899	0.463	0.874
F-Score	0.762	0.844	0.596	0.791

Figure 5: Scores by types and POS tag

To evaluate type accuracy, we use a measure similar to the Jaccard distance between true and induced inflectional forms for a lemma. Unlike problems of soft clustering, it is possible to define what is a correct clustering in a lemmatization problem. The precision of an individual clustering can be defined as the size of the intersection between an induced set of inflectional forms and the standard set of inflectional forms divided by the size of the standard set.

	LEMMA	INFLECTIONS
VERBS	*ergänzen*	unternommenen betriebsrat ergänzung abrunden abkehr zusammenführen vervollständigen weswegen flüchtlingskonvention entwicklungschancen staatsangehörigkeit ergänzend ergänzen einander durchschlagen . . .
	sterben	sterben verhungern helfern designierten jährlich zutritt meistens amerikanern irakern fünfte tod planeten industriegebieten fonds dramatisch us-regierung
NOUNS	*knie*	asiatischen zusammengestellt zufügt kniefall knie knien apartheid-regime rechtsanspruch
	euro	ausübt euroraums euroumstellung euro-raums euros euro-länder euro-ländern eurozusammenarbeit euro euro-raum euroländer euro-system . . .

Figure 3: A list of German candidate verb and noun lemmas and their inflected forms extracted automatically through alignment and transitive linkage. List of candidate inflections is unordered either in terms of frequency or in terms of dictionary precedence.

By summing the individual clustering precision figures over the entire set Λ of sets of inflectional forms I_i, and normalizing this by $N = |\Lambda|$ the precision is calculated as

$$\frac{1}{N} \sum_{I_i \in \Lambda} \frac{|I_i \cap I_g|}{|I_i|}$$

Similarly, recall is defined similar to the above but divided by $|I_g|$ instead:

$$\frac{1}{N} \sum_{I_i \in \Lambda} \frac{|I_i \cap I_g|}{|I_g|}$$

These results are given in Fig. 5.

6 Conclusion

We have outlined a minimally supervised approach to inducing a lemmatization scheme for a target language using alignment and transfer methods across parallel bilingual corpora. Compared to the few previous studies on lemmatization (Yarowsky and Wicentowski, 2000; Yarowsky et al., 2001), we have reduced the level of supervision necessary to a bare minimum, obviating any need to manually select a set of "dictionary entry forms" for the target language, while retaining a time complexity that is feasible in spite of a lack of predefined assumptions, even when the parallel corpora span some 25M words for both source and target language with app. 200K word types in the target text.

In future studies, to further increase the robustness and accuracy of the approach, several avenues of investigation will have to be included. First, given that it is possible to automatically generate a POS tagger (its robustness and accuracy notwithstanding) for a target language through alignment and transfer methods, it should be possible to leverage such additional information to enhance the accuracy and coverage of our lemmatization method. Second, given current developments in the field, it would be possible to generalize over the induced lemmata set to generate new inflections. To do so would require induction of abstract inflectional patterns in the target language for what may or may not be equivalent or analogous to number, case, tense, mood, voice, etc. which would require the incorporation of all the lemmata over all POS tags observed in English (e.g. prepositions, pronouns, conjunctions, etc.) as well as the syntactic information generated by parsers.

Acknowledgements

This work was supported by NSF grant BCS-0651988. The authors would also like to thank Jason Baldridge and Alexis Palmer for providing invaluable comments on the paper.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604,

Morristown, NJ, USA. Association for Computational Linguistics.

Sabine Brants and Silvia Hansen. 2002. Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pages 1643–1649, Las Palmas.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1):3.

James R Curran and Stephen Clark. 2003. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*.

Mona Diab. 2000. An unsupervised method for multilingual word sense tagging using parallel corpora: a preliminary investigation. In *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality*, pages 1–9, Morristown, NJ, USA. Association for Computational Linguistics.

Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 49–56, Ann Arbor, Michigan, June. Association for Computational Linguistics.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguististics*, 27(2):153–198.

Sharon Goldwater. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Harald Hammarström. 2006. A naive theory of affixation and an algorithm for extraction. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 79–88, New York City, USA, June. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Nat. Lang. Eng.*, 7(3):207–223.

Taesun Moon and Jason Baldridge. 2007. Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 390–399.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sylwia Ozdowska. 2006. Projecting POS tags and syntactic dependencies from English and French to Polish in aligned corpora. In *EACL 2006 Workshop on Cross-Language Knowledge Induction*.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 102–109, Morristown, NJ, USA. Association for Computational Linguistics.

Manabu Sassano. 2001. An empirical study of active learning with support vector machines for Japanese word segmentation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 505–512, Morristown, NJ, USA. Association for Computational Linguistics.

Jung H. Shin, Young S. Han, and Key-Sun Choi. 1996. Bilingual knowledge acquisition from Korean-English parallel corpus using alignment method: Korean-English alignment at word and phrase level. In *Proceedings of the 16th conference on Computational linguistics*, pages 230–235, Morristown, NJ, USA. Association for Computational Linguistics.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 207–216, Morristown, NJ, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT '01: Proceedings of the first international conference on Human language technology research*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

The Unstructured Information Management Architecture: Towards an Interoperability Standard for Text and Multi-Modal Analytics

Eric Nyberg
Carnegie Mellon University
ehn@cs.cmu.edu

Unstructured information is typically the direct product of human communications. Examples include natural language documents, email, speech, images and video. For unstructured information to be processed by traditional applications, it must first be analyzed to assign application-specific semantics to the unstructured content. An example of assigning semantics includes identifying and wrapping regions of text in a document with XML tags which identify places where persons, organizations, times, or events are mentioned.

Recent growth in unstructured information management (UIM) applications is largely driven by the wealth of unstructured information found on the external web, in corporate intranets, document repositories, call-centers, and in customer and employee business communications. UIM applications tend to be highly decomposable; that is, they may be broken down into finer grained parts, each performing a specialized function in an overall analysis workflow. Each of these functions, or *analytics*, may be reused in different workflows to perform different aggregate analyses. Because UIM applications are naturally decomposable and reusable across different solutions, there is an advantage to defining an architecture that supports standards for interoperability between analytic functions.

The *Unstructured Information Management Architecture* (UIMA) refers to a software architecture for defining and composing interoperable text and multimodal analytics. UIMA builds on prior work at IBM on frameworks for text and multimodal analytics, including TAF, TALENT and WebFountain. It has been inspired and influenced by other projects, including TIPSTER, Mallet, GATE, OpenNLP, Atlas and Catalyst. In late 2004 IBM released the *UIMA Software Developers Kit (SDK)* on IBM alphaWorks[1]. The SDK is freely available and provides the tools and run-time necessary for creating, composing and deploying component analytics. Since 2004 industrial, academic and government research & development projects have applied the UIMA SDK as a foundation for building and/or enhancing applications that process unstructured information.

In early 2006 IBM contributed an implementation of the UIMA Framework to the open-source community through Source Forge. In 4Q 2006, the open-source framework was accepted as an Apache incubation project, where IBM and non-IBM committers participate in its collaborative development. Any given framework implementation, however, may not satisfy the requirements of all UIM applications. For example, some applications may require very lightweight browser-based analytics, while others may require heavyweight, carefully managed, highly scalable solutions. The diverse variety of potential implementations suggests that there should be a more general specification for interoperability that may allow for different framework implementations and different levels of compliance, facilitating interoperability for a broader range of application and programming requirements.

To help define a broader, platform independent standard that can guide the open-source development of Apache UIMA and other related frameworks, a Technical Committee was formed in late 2006 to develop a standard specification under the auspices of OASIS[2], a

[1] http://www.ibm.com/alphaworks/tech/uima
[2] http://www.oasis-open.org

Standards Development Organization (SDO). The intent is that such a standard will allow different frameworks to emerge, while also allowing applications built on different platforms and programming models to have a standard means to share analysis data and analytic services.

This presentation summarizes the elements of the UIM architecture specification, which are aligned with the specific interoperability requirements of UIM applications:

- **Data Representation**. Support the common representation of artifacts and artifact metadata (analysis results) independently of artifact modality and domain model.

- **Data Modeling and Interchange**. Support the platform-independent interchange of analysis data in a form that facilitates a formal modeling approach and alignment with existing programming systems and standards.

- **Discovery, Reuse and Composition**. Support the discovery, reuse and composition of independently-developed analytics.

- **Service-Level Interoperability**. Support concrete interoperability of independently developed analytics based on a common service description and associated SOAP bindings.

Elements of the specification will be illustrated with examples from current UIM applications, followed by a discussion of the current status of the specification and open issues to be addressed in ongoing committee work.

Reference

Ferrucci, D., A. Lally, D. Gruhl, E. Epstein, M. Schor, J. W. Murdock, A. Frenkiel, E. W. Brown, T. Hampp, Y. Doganata, C. Welty, L. Amini, G. Kofman, L. Kozakov and Y. Mass (2006). “Towards an Interoperability Standard for Text and Multi-Modal Analytics”, IBM Research Report RC24122 (W0611-188), November.

Towards a Uniform Representation of Treebanks: Providing Interoperability for Dependency Tree Data

Olga Pustylnikov
Department of
Computational Linguistics and
Texttechnology
Bielefeld University
D-33615 Bielefeld, Germany
Olga.Pustylnikov@uni-bielefeld.de

Alexander Mehler
Department of
Computational Linguistics and
Texttechnology
Bielefeld University
D-33615 Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Abstract

In this paper we present a corpus representation format which unifies the representation of a wide range of dependency treebanks within a single model. This approach provides interoperability and reusability of annotated syntactic data which in turn extends its applicability within various research contexts. We demonstrate our approach by means of dependency treebanks of 11 languages. Further, we perform a comparative quantitative analysis of these treebanks in order to demonstrate the interoperability of our approach.

1 Introduction

In recent years a large number of natural language resources providing structured information by means of annotated data have been developed. Among them, different types of corpora consisting, for example, of texts, multimodal documents, web documents or syntactic treebanks serve different scientific purposes and are available by means of specific schemata. To refer to these different types we speak of *corpus genres*.

Treebanks, instantiating a specific corpus genre, are syntactically annotated corpora which are mainly used in data oriented approaches to computational linguistics (Bod et al., 2003). The availability of these corpora is crucial for training and testing NLP applications as well as for exploring linguistic phenomena. Fortunately, a large number of syntactic treebanks is available for a multitude of languages.[1] However, these treebanks are provided in a wide range of different formats. That is, NLP tools as, e.g., syntactic parsers which are trained on a variety of languages in order to provide cross-lingual interoperability have to be adapted to ever new representation formats of such banks. Thus, a major problem of using treebanks in NLP is the high effort of adaptating the tools or transforming into the formats. A unification of existing formats reduces this effort and makes different treebanks applicable to divergent tools via a single interface. Although reusability of treebanks has a high priority (Kakkonen, 2005), mapping them onto a single format is not an easy task. This is explained with respect to three levels of corpus related features:

- **Level 1** refers to corpus genre related features.
- **Level 2** relates to specifics of the object data.
- **Level 3** includes features induced by the operative representation format.

On level 1 we distinguish, for example, between *dependency* and *constituency* structure-related treebanks. This distinction reflects different syntactic theories underlying the generation of the treebanks. Focusing on the corpus genre of dependency treebanks, they can be further distinguished with respect to the annotation requirements induced by the target language or by the specific dependency grammar in use.[2] This happens on level 2. On level 3 we distinguish formats of the target treebank (as, e.g., the

[1] See (Kakkonen, 2005) for a review on existing treebanks.

[2] See (Nivre, 2005) for a review on different dependency grammars.

Penn Treebank (Marcus et al., 1993), TUT (Bosco et al., 2000), NEGRA (Skut et al., 1998) or SUSANNE (Sampson, 1995))

In this paper we transform dependency treebanks of 11 languages into a single format in order to provide interoperability of cross-lingual NLP systems operating on them. For this task we take level 1, i.e. corpus genre-related, and level 3, i.e. format related differences into account. Thus, we present a format general enough to map dependency *and* constituency structures, but concentrate on dependency treebanks whose level 3 differences are eliminated. Note that we do not consider differences induced by the object data, that is level 2 features. The reason is that their elimination is more difficult (as in the case of dependency grammar-related differences) or even impossible (as in the case of language specific features). In summary, the present paper overcomes the deficit of a lacking representation format which maps the existing variety of treebanks and, thus, provides interoperability on the level of syntactic ontologies. We demonstrate this interoperability by a quantitative structure analysis, which – to the best of our knowledge – is the first one operating on 11 languages. Note that all freely available corpora being analyzed in this study can be downloaded from our web site.[3]

The paper is organized as follows: Its conceptual framework is described in Sec. 2. Sec. 3 presents our experimental setting which benefits from the unified representation of dependency treebanks. Sec. 4 presents the results of our comparative study. Finally, Sec. 5 discusses our findings while Sec. 6 gives a conclusion and prospects future work.

2 Towards a Unified Representation for Treebanks

Treebanks may differ with respect to genre, data, or format-related criteria as discussed in Sec. 1. It is highly desirable to reduce this variety in order to achieve better data access and reusability (Kakkonen, 2005). The starting point for providing portability of corpus data is to use XML as the primary format of data exchange. In the past, specific XML models were provided for representing instances of specific corpus genres, languages (and sometimes even of specific linguistic theories). An example is TIGER-XML which optimally fits syntactic treebanks but is not applicable to other types of corpora. In contrast to this, what we search for is a *generic* format allowing

- to integrate all kinds of treebanks and
- to be extensible to map the specifics of a given language or grammar.

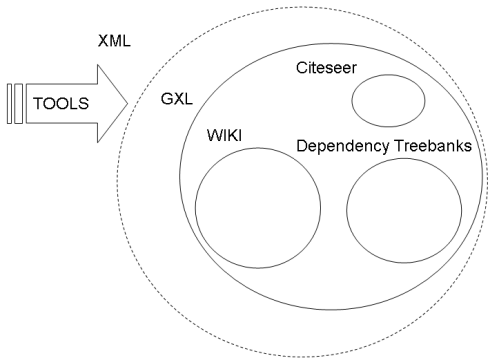


Figure 1: The Scope of GXL for corpus representation.

GXL (Holt et al., 2006) (see Figure 1) is an XML-based graph representation format which satisfies these two requirements as it allows one to deal with all kinds of graph structures.[4] Due to its generic graph data model it was successfully applied to various corpus genres such as Wikipedia-based corpora, newspaper corpora or dependency treebanks.[5] GXL represents corpus units as `node`-elements and their relations as `rel`- (or `edge`-) elements. All additional information of nodes (as, e.g., POS information) is stored within `attr`-elements. Each node is given an `id` so that it can be accessed via IDREF[6] attributes stored in `relend`-elements (in terms of the `target`- attribute – cf. Figure 2). The `direction` attributes (`in`/`out`) describe the kind of relationship between the nodes as, e.g., a head(`in`)-modifier(`out`) relation in a dependency tree.

[3] `http://ariadne.coli.uni-bielefeld.de/indogram/resources/`.

[4] See (Diestel, 2005) for the definition of a graph.

[5] See e.g. (Mehler and Gleim, 2005), (Mehler et al., 2007), (Ferrer i Cancho et al., 2007).

[6] IDentifier REFerence is a reference to a unique identifier in XML.

This general model allows the representation of various types of corpora and to store an arbitrarily large amount of additional information in order to account for their differences. That is, by mapping treebanks onto a GXL-based data model we enable tools to operate on different treebanks using the same interface and at the same time preserve their differences using a single representation format.

```
<rel>
  <relend direction="in" target="s30_7" />
  <relend direction="out" target="s30_8" />
</rel>
```

Figure 2: A dependency link between two nodes.

However, it turns out that this goal cannot be provided by GXL straightforwardly, but only by an extension of it henceforth called e*xtended* GXL (eGXL). The reason is that in many cases it is desirable to extend GXL to achieve a better fit to the object data. Due to its high level of generality, eGXL is less compact at some points. Consider, for example, the GXL-based representation of a token of the SUSANNE corpus (Sampson, 1995):

```
<node id="J02_p1_11">
    <attr name="lemma">
        <string>make</string>
    </attr>
    <attr name="pos">
        <string>VVNv</string>
    </attr>
    <attr name="word">
        <string>made</string>
    </attr>
</node>
```

Using GXL the token "made" is represented by 11 lines (3 lines per attribute). This is verbose from the point of view of data storage and retrieval. The specification of `attr` elements (`<string>...</string>`) is required from the schema allowing to store different data types like `string`, `integer`, etc. A more compact representation comparable to TIGER-XML encodes all extra information of a node by means of attributes in a single line. Thus, it should be possible to encode tokens from the SUSANNE corpus as follows:

```
<node id=".." form=".." lemma=".." />
```

One solution to achieve this is to extend GXL by means of additional node attributes providing a more compact storage of the data. Unfortunately, the original GXL-Schema[7] contains syntax errors making it impossible to derive from it. Thus, in a first step we corrected the errors[8] and extended the model by two `node` attributes, namely `form` and `lemma`[9]. Further, we added an attribute (`type`) to the `relend`-element to specify the type of relation. It allows one to define different types of relations once in the head of the document and access them by means of reference attributes.

What we get is a new graph representation scheme extendedGXL (eGXL)[10] which integrates the possibilities of TIGER-XML to represent syntactic trees. eGXL is extensible and can be adapted to more specific data while it remains generic being applicable to any kind of corpora.

```
<node id="Types">
  <graph id="g0">
    <node id="POS"/>
    <node id="t1" name="verb"/>
    <node id="t3" name="prepozitie"/>
    <node id="t5" name="substantiv"/>
    ...
    <node id="CAT"/>
    <node id="t2" name="subiect"/>
    <node id="t4" name="atribut subst."/>
    ...
    <edge from="POS" to="t1"/>
    <edge from="CAT" to="t2"/>
    <edge from="POS" to="t3"/>
    ...
  </graph>
</node>
```

Figure 3: eGXL *Types* graph.

```
<node id="Sentences">
   <graph id="g1">
     <node id="s1_1" form="Autorizatia" pos="t1" cat="t2"/>
     <node id="s1_2" form="pentru" pos="t3" cat="t4"/>
       ...
     <rel>
       <relend direction="in" target="s1_7"/>
       <relend direction="out" target="s1_1"/>
     </rel>
     <rel>
       <relend direction="in" target="s1_1"/>
       <relend direction="out" target="s1_2"/>
     </rel>
       ...
   </graph>
```

Figure 4: eGXL *Sentences* graph.

[7] `http://www.gupro.de/GXL/xmlschema/gxl-1.0.xsd`.

[8] See (Pustylnikov, 2007b) for details on error removal.

[9] Which seem to map important pieces of information since we observed them in almost all treebanks.

[10] `http://ariadne.coli.uni-bielefeld.de/indogram/resources/XML\%20Schemata/eGXL-1.0.xsd`.

```
1   Cathy   Cathy   N   N   eigen|ev|neut   2   su   _   _
2   zag zie V   V   trans|ovt|1of2of3|ev    0   ROOT    _   _
3   hen hen Pron    Pron    per|3|mv|datofacc   2   obj1    _   _
4   wild    wild    Adj Adj attr|stell|onverv   5   mod _   _
5   zwaaien zwaai   N   N   soort|mv|neut   2   vc   _   _
6   .   .   Punc    Punc    punt    5   punct   _   _

<sentence id="8" user="" date="">
 <word id="1" form="Detta" postag="POOP" head="2" deprel="OO"/>
 <word id="2" form="vill" postag="WVPS" head="0" deprel="ROOT"/>
 <word id="3" form="jag" postag="POPPHH" head="2" deprel="SS"/>
 <word id="4" form="bestämt" postag="AJ" head="2" deprel="AA"/>
 <word id="5" form="bemöta" postag="VVIV" head="2" deprel="VG"/>
 <word id="6" form="." postag="IP" head="2" deprel="IP"/>
</sentence>

************** FRASE ALB-2 **************
1 Valona (VALONA NOUN PROPER F Â$CITY) [1.10;VERB-SUBJ]
1.10 t [] (ESSERE VERB MAIN IND PRES INTRANS 3 SING) [0;TOP-VERB]
2 in (IN_MANO_A PREP POLI LOCUTION) [1.10;VERB-PREDCOMPL+SUBJ]
3 mano (IN_MANO_A PREP POLI LOCUTION) [2;CONTIN+LOCUT]
4 ai (IN_MANO_A PREP POLI LOCUTION) [3;CONTIN+LOCUT]
4.1 ai (IL ART DEF M PL) [2;PREP-ARG]
5 dimostranti (DIMOSTRANTE NOUN COMMON ALLVAL PL) [4.1;DET+DEF-ARG]
6 . (#\. PUNCT) [1.10;END]
```

Figure 5: 3 Treebanks: Dutch, Swedish and Italian.

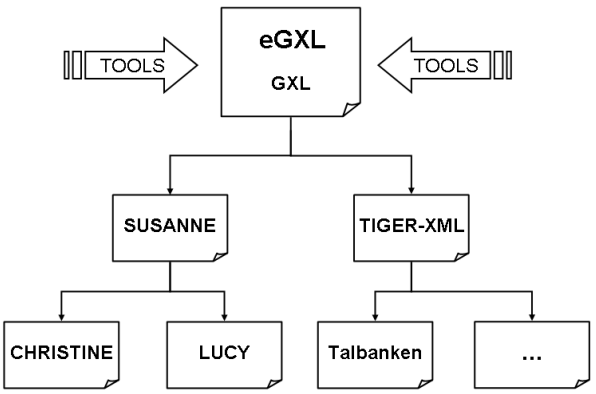


Figure 6: eGXL Hierarchy

2.1 The structure of eGXL

Figures 3 and 4 illustrate the general structure of eGXL. An eGXL document consists of two logical parts. The first part is the `Types` graph containing all attributes with the corresponding `ids`. That is, all possible values of an attribute (e.g. POS) are listed only once at the beginning of the document and accessed later by reference attributes. The second part consists of graph-based representations of sentences where the words are represented as `nodes` and their dependency relations as `rels`.

Generally speaking, treebanks vary with respect to content-related attributes (e.g., the `POS` attribute). Other than the `form` and `lemma` attributes, they are not part of the basic schema of eGXL. In order to map corpus specifics induced by such attributes we

```
<graph id="g1">
    <node id="s0_1" form="Cathy" lemma="Cathy" pos="t1" .../>
    <node id="s0_0"/>
    <node id="s0_2" form="zag" lemma="zie" pos="t4" extra="t4" ... />
    <node id="s0_3" form="hen" lemma="hen" pos="t7" extra="t7" ... />
    <node id="s0_4" form="wild" lemma="wild" pos="t10" .../>
    <node id="s0_5" form="zwaaien" lemma="zwaai" pos="t1" .../>
    <node id="s0_6" form="." lemma="." pos="t15" extra="t15" ...>
<rel>

<graph id="g8">
    <node id="s8_1" form="Detta" pos="t151" cat="t298"/>
    <node id="s8_2" form="vill" pos="t245" cat="t187"/>
    <node id="s8_0"/>
    <node id="s8_3" form="jag" pos="t152" cat="t306"/>
    <node id="s8_4" form="bestamt" pos="t26" cat="t254"/>
    <node id="s8_5" form="bemöta" pos="t227" cat="t312"/>
    <node id="s8_6" form="." pos="t86" cat="t86"/>
<rel>

<graph id="g2">
   <node id="n2_1" form="Valona" lemma="VALONA">
     <graph id="gn2_1">
       <edge from="n2_1" to="t16"/>
       <edge from="n2_1" to="t47"/>
       <edge from="n2_1" to="t7"/>
       <edge from="n2_1" to="t48"/>
     </graph>
   </node>
   <node id="n2_0" form="root"/>
   <node id="n2_1.10" form="t" lemma="ESSERE" />

<rel>
```

Figure 7: 3 Treebanks in eGXL: Dutch, Swedish and Italian.

provide a mechanism of extending eGXL which induces a family of XML Schemata all being derived from the basic eGXL schema (see Figure 6).

This model is not restricted to a particular treebank since all specific information of a treebank is instantiated by the `Types` graph. The core structure of the document remains always the same allowing the document to be accessed with the same tools. Figures 5 and 7 illustrate how three different input formats are transformed into a single representation. Figure 7 contains the sentences from Figure 5 transformed into eGXL. Although the sentences originate from different source treebanks they can be treated as a part of a single document regarding their structure.[11]

3 Quantitative Profiling of Dependency Treebanks

The unified representation format for dependency treebanks provided by eGXL allows us to compare 11 languages (Table 1) according to their quantitative characteristics. We benefit from our unified rep-

[11] Since the Italian treebank originally contains unlabeled attributes we include an additional attribute graph into a node element (Figure 7, the lowest part). This representation expands the node element, but the main shape of the document is preserved illustrating the extensibility of eGXL.

Treebank	Language	Size (#token)	Reference
Alpino Treebank v. 1.2	Dutch	195.069	(van der Beek et al., 2002)
Danish Dependency Treebank v. 1.0	Danish	100.008	(Kromann, 2003)
Sample of sentences of the Dependency Grammar Annotator	Romanian	36.150	`http://www.phobos.ro/roric/DGA/dga.html`
Russian National Corpus	Russian	253.734	(Boguslavsky et al., 2002)
A sample of the Slovene Dependency Treebank v. 0.4	Slovene	36.554	(Džeroski et al., 2006)
Talkbanken05 v. 1.1	Swedish	342.170	(Nivre et al., 2006)
Turin University Treebank v. 0.1	Italian	44.721	(Bosco et al., 2000)
CESS - Catalan Dependency Treebank	Catalan	100.000	(Civit et al., 2004)
Cast3LB - Spanish Dependency Treebank	Spanish	100.000	(Civit and Martí, 2005)
Prague Dependency Treebank 2.0	Czech	1.957.247	(Hajič, 1998)
BulTreeBank	Bulgarian	196.000	(Osenova and Simov, 2004)

Table 1: General Properties of the Treebanks.

resentation which provides a maximal reduction of level 3 differences (Sec. 1). The quantitative characteristics relate to dependency trees. The idea to compare languages by means of such features stems from (Ferrer i Cancho et al., 2004) who transformed 3 treebanks into Global Syntactic Dependency Networks (GSDNs) in order to measure their similarities. The nodes of GSDNs model are tokens where edges occur between two nodes if there is at least one dependency link between the corresponding tokens in the input bank. (Ferrer i Cancho et al., 2007) found out that the GSDNs of seven languages exhibit similar network properties which seem to be possibly universal properties of these kinds of networks.

Obviously, treebanks cannot be distinguished in terms of such measures. Thus, we focus on a different set of their structural characteristics. The aim is to answer the following questions:

- Can we classify treebanks by means of quantitative properties?
- Does the explored classification relate to known differences of the languages being analyzed?

In summary, we treat the above questions as a classification task in terms of *quantitative structure analysis* (Pustylnikov, 2007a; Mehler et al., 2007) using feature vectors to represent structural properties of treebanks.

3.1 Quantitative Dependency Tree Characteristics

We treat dependency trees of sentences as the basic unit to compute the characteristics listed below for each of the 11 input corpora. In order to get a single value of these characteristics for each of the corpora we average over all sentence-related observations of the respective corpus. The quantitative characteristics being computed are defined as follows:

In and Out Degree: The in (out) degree is given by the number of outgoing (incoming) dependency links observed for each word in the corpus.

Sentence Length: The sentence length is the average sentence length of a treebank.

Depth: The depth is the average depth of a dependency tree (sentence).

Depth Imbalance: As a measure of the imbalance of the sentence trees of a treebank we compute their A*bsolute* D*epth* I*mbalance* (ADI) according to (Botafogo et al., 1992). Starting from an input vertex v, this measure basically computes the standard deviation of the adjusted heights of v's child nodes. We compute the ADI for the root vertex r of the sentence tree T of each dependency treebank, where the higher $\mathrm{ADI}(r) \in \mathbb{R}_+$ the higher the variance among

the heights of r's child nodes, the more imbalanced T.

Child Imbalance: By analogy to the ADI we also compute the A*bsolute* C*hild* I*mbalance* (ACI) (Botafogo et al., 1992). Whereas the ADI evaluates imbalance in terms of the heights of child nodes, the ACI focuses on the sizes of the trees dominated by these nodes. Size is measured as the number of vertices of the respective tree. Obviously, the ADI also reflects the width of a tree and, thus, provides complementary information to the ACI.

Compactness and Stratum: Finally, the stratum and the compactness measures – as introduced by (Botafogo et al., 1992) – operate on graphs. In the present case we apply the measures to sentence trees which can be described as subsets of graphs. The Stra*tum* (Stra) is a metric which measures, so to speak, the deviation of a given sentence graph (tree) from a purely linearly organized graph with the same number of vertices where a stratum of 1 indicates a maximally hierarchically organized sentence. The C*ompactness* (C) analogously varies from 0 (i.e. graphs that are completely disconnected) to 1 (i.e. graphs that correspond to completely connected graphs). In our case the maximal values of 0 and 1 are never achieved since we deal with trees, which in turn are never completely connected or disconnected. Nevertheless, we expect the (C) values to vary for the different dependency treebanks reflecting different sentence structures.

3.2 Quantitative Structure Analysis

In text classification, structural features revealed to be a good alternative to the traditional *bag of words* approach (Pustylnikov and Mehler, 2007; Mehler et al., 2007). To build the feature vectors for our language-related classification task we compute the values of the characteristics listed in Sec. 3.1 for the 11 treebanks after being transformed into the eGXL. The aggregation of the feature values was done by computing the mean, the standard deviation and the entropy of the corresponding corpus-related value distributions. Each treebank is finally represented by a numerical vector with the cardinality $M \times N$ where M represents the number of characteristics and $N = 3$ is the number of location and dispersion parameters in use. To classify the treebanks we use semi-supervised hierarchical clustering. The results obtained in the experiment are presented in Sec. 4.

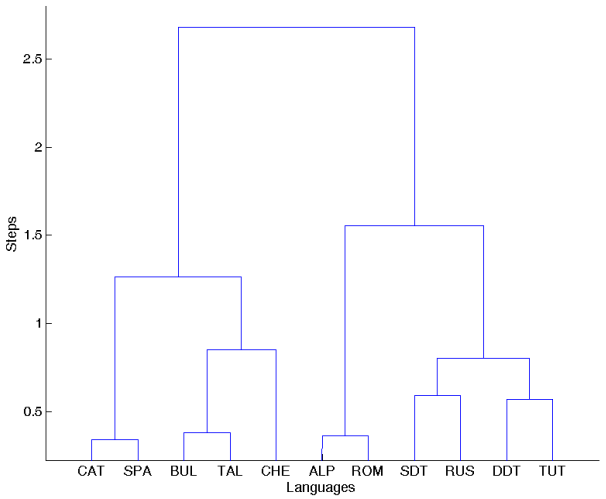


Figure 8: Results from language clustering. SDT - Slovenian, ROM - Romanian, RUS - Russian, DDT - Danish, ALP - Dutch, SPA - Spanish, TUT - Italian, BUL - Bulgarian, TAL - Swedish, CAT - Catalan, CHE - Czech.

4 Experimental Results

In Figure 4 the clustering results are visualized by a dendrogram. Each associated pair expresses the maximal similarity between two clusters among all clusters in every step. Thus, on the bottom the most similar languages according to the treebank characteristics are combined. The similarity threshold increases iteratively so that less similar clusters become connected. Finally, all groups constitute a single cluster (*bottom-up* clustering).[12] The height of the connection between two clusters indicates the strength of the similarity between them, that means, the higher the connection, the less similar are the two clusters to each other (Manning and Schütze, 1999).

In our case, Dutch (ALP) and Romanian (ROM) or Spanish (SPA) and Catalan (CAT) exhibit the

[12] Alternatively we can start from one cluster including all languages and divide them by reducing the similarity threshold in every iteration step (*top-down* clustering).

greatest similarity (the lowest connections) whereas the least similarity is expressed by means of the highest connection combining all languages within one cluster. A detailed discussion of the findings is presented in Section 5.

5 Discussion

At first glance, the overall partition of languages into clusters is far from their genetic classification (i.g. germanic, romance, slavic etc.). Looking at the similarity threshold of around 1.5 in the middle of the dendrogram we can point out three clusters differing internally in size and in similarity degrees. The first cluster contains SPA and CAT which (together with ROM and ALP) have the lowest connection as well as Bulgarian (BUL), Swedish (TAL) and Czech (CHE). The second cluster consists of two languages: ROM and ALP which are also dissimilar to other languages since they become merged with the third cluster only around the threshold of 1.6. The third cluster combines Slovenian (SDT) and Russian (RUS) and Danish (DDT) and Italian (TUT). The close connection between RUS and SDT is in accordance with our intuition about the membership of both languages in the slavic family. Similarly, SPA and CAT which are closely related genetically are also grouped together. Connections between languages which cannot be attributed to their genetic relationships require other explanations. Obviously, languages which differ genetically can nevertheless share structural properties (e.g. with respect to syntax or morphology). Since the observations we make about languages are related to dependency structures there are many possible reasons letting the sentence structure exhibit a particular shape. Italian for example is a Romance language which has preserved its inflectional morphology which may relate it to Slavic languages with respect to its structure. Bulgarian, a South-Slavic language has a tendency towards isolating / agglutinating languages which makes it group together with Swedish (TAL) and Czech (CHE). To complete the picture and to verify the assumptions further investigations need to be carried out which are beyond the scope of the present study. Here, we aimed at illustrating the possibilities for comparative investigations which arise with a unification of corpora.

6 Conclusion

In this paper we introduced a new XML based format for treebank representation. The format corrects and extends the generic graph model GXL to provide an effective means of integrating additional information in terms of `node` attributes. Our main goals have been

a) to provide a unification of 11 dependency treebanks,

b) to develop a representation format allowing to integrate the peculiarities of particular treebanks and

c) to combine the benefits of existing formats like TIGER-XML, etc. within a single representation.

We illustrated the potential of eGXL-based dependency treebank representations by a quantitative study. A unification of treebanks developed under different conditions is a demanding task with respect to format, language and annotation specific differences. A unification on the level of format by means of eGXL enabled the comparative quantitative investigation of 11 languages with a elevenfold reduction in computation effort. The results are in part interpretable in terms of language typology as in case of Slovene and Russian as well as of Spanish and Catalan. A systematic study of the impact of quantitative characteristics of dependency trees on language classification will be part of future work.

References

Rens Bod, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. CSLI Publications, Stanford.

Igor Boguslavsky, Ivan Chardin, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreidlin, and Nadezhda Frid. 2002. Development of a dependency treebank for russian and its possible applications in NLP. In *Proc of LREC 2002*.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Lesmo Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proc. of LREC 2000*.

Rodrigo A. Botafogo, E. Rivlin, and B. Shneiderman. 1992. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

Montserrat Civit and M.A. Martí. 2005. Building Cast3LB: A Spanish Treebank, a Research on Language and Computation. *Springer Verlag*, pages 549–574.

M. Civit, N. Bufí i, and P. Valverde. 2004. CAT3LB: a Treebank for Catalan with Word Sense Annotation. In *TLT2004*, pages 27–38. Tubingen University.

Reinhard Diestel. 2005. *Graph Theory*. Springer, Heidelberg.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtský, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of LREC 2006*.

Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69:051915.

Ramon Ferrer i Cancho, Alexander Mehler, Olga Pustylnikov, and Albert Díaz-Guilera. 2007. Correlations in the organization of large-scale syntactic dependency networks. In *TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pages 65–72.

Jan Hajič. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In E. Hajičová, editor, *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová*, pages 106–132. Karolinum, Charles University Press, Prague, Czech Republic.

Richard C. Holt, Andy Schürr, Susan Elliott Sim, and Andreas Winter. 2006. GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170.

Tuomo Kakkonen. 2005. Dependency Treebanks: Methods, Annotation Schemes and Tools. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 94–104, Joensuu, Finland.

Matthias T. Kromann. 2003. The danish dependency treebank and the underlying linguistic theory. In Joakim Nivre and Erhard Hinrichs, editors, *Proc. of TLT 2003*. Växjö University Press.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. In *Computational Linguistics 19*.

Alexander Mehler and Rüdiger Gleim. 2005. The net for the graphs – towards webgenre representation for corpus linguistic studies. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as corpus*, pages 191–224. Gedit, Bologna, Italy.

Alexander Mehler, Peter Geibel, and Olga Pustylnikov. 2007. Structural Classifiers of Text Types: Towards a Novel Model of Text Representation. *To appear in: LDV Forum*.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proc. of LREC 2006*.

Joakim Nivre. 2005. Dependency Grammar and Dependency Parsing. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.

Petya Osenova and Kiril Simov. 2004. BTB-TR05: BulTreeBank Stylebook. BulTreeBank Project Technical Report Nr. 05. Technical report, Linguistic Modelling Laboratory, Bulgarian Academy of Sciences.

Olga Pustylnikov and Alexander Mehler. 2007. Structural differentiae of text types. a quantitative model. In *Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl)*.

Olga Pustylnikov. 2007a. Guessing Text Type by Structure. In *Proceedings of the ESSLLI Student Session '07*, pages 221–231.

Olga Pustylnikov. 2007b. GXL-extension (GXXL-1.0). Correcting errors in GXL.

Geoffrey Sampson. 1995. *English for the Computer*. Clarendon Press, Oxford.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.

Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. 2002. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands CLIN*, Radopi.

Linking and Integrating two Electronic Lexicons

Nilda Ruimy, Adriana Roventini, Rita Marinelli, Marisa Ulivieri
Istituto di Linguistica Computazionale – CNR
Via Moruzzi,1 – 56124 – Pisa, Italy
{nilda.ruimy,adriana.roventini,rita.marinelli, marisa.ulivieri}
@ilc.cnr.it

Abstract

Nowadays, in the field of Computational Lexicography, much attention is being paid, when building lexical resources, to their interoperability and their easy integration in HLT-NLP applications for an enhanced performance. Concerning already existing computational lexicons, on the other hand, their integration and interoperability is attainable, provided their main features offer a field of comparison. The two largest and extensively encoded electronic lexicons of Italian language fulfill this essential requirement. Although developed according to two different lexical models, ItalWordNet and PAROLE-SIMPLE-CLIPS present in fact many compatible aspects. Linking and eventually merging these lexical resources in a common representation framework seems therefore a wise move to offer the end-user a more exhaustive and in-depth lexical information combining the potentialities and most outstanding features offered by the two lexical models. This paper reports on the ongoing linking of the two lexicons. The mapping of the ontologies on which basis the lexicons are structured is described; an overview of the adopted methodology, of the linking process and of the results of the first mapping phase regarding 1st Order Entities is provided. Reciprocal benefits and enhancements for the two resources are also illustrated that definitely justify the soundness of our linking initiative.

1 Introduction

'As the need for cross-lingual studies and applications grows, it is increasingly important to develop resources in the world's languages that can be compared and linked, used and analyzed with common software, and that contain linguistic information for the same or comparable phenomena.'[1].

Nowadays, in the field of Computational Lexicography, much attention is being paid, when building lexical resources, to their interoperability and their easy integration in HLT-NLP applications for an enhanced performance. The most relevant collaborative efforts that lexicon experts devoted to developing consensual specifications and enforcing standards in this domain have led to the creation of the Lexical Markup Framework (LMF)[2], a metamodel which provides a common standardized framework for the construction of computational lexicons.

Concerning already existing electronic lexicons, on the other hand, their integration and interoperability is attainable, provided their main features offer points of comparison. The two largest and extensively encoded lexicons of Italian language, which were developed during the last decade at the CNR Institute of Computational Linguistics in Pisa, fulfill this essential requirement.

[1] First International Conference on Global Interoperability for Language Resources ICGL2008, Call for Papers, Mission.

[2] in the International Organization for Standardization (ISO) sub-group TC37/SC4/WG4.

2 Lexical Resources

ItalWordNet[3] (henceforth IWN) is a lexical semantic database created in the framework of the *EuroWordNet* (EWN) project[4] and extended in the national project *Integrated System for the Automatic Language Treatment* (SI-TAL). It is based on the EuroWordNet lexical model[5] (Vossen, 1998) which is, in turn, inspired to the Princeton WordNet (Miller *et al.*, 1990).

IWN (Roventini *et al.*, 2003) provides the semantic description of 67,000 Italian word senses (verbs, common and proper nouns, adjectives, adverbs and multi-word units), which are clustered in about 50,000 *synsets* (i.e. synonym sets). One of the salient features of the resource is the connection of all IWN synsets to the Princeton Wordnet database (Fellbaum, 1998). Such synsets, that represent lexicalized concepts, are classified in terms of an ontology and interconnected by means of a set of semantic relations that link both intracategorial and intercategorial synsets (Alonge *et al.* 1998).

The IWN Top Ontology (henceforth, TO), which slightly differs from the EWN TO[6], is a hierarchy of 65 language-independent Top Concepts (TCs) clustered in three main categories distinguishing 1stOrderEntities, 2ndOrderEntities and 3rdOrder Entities. Their subclasses, hierarchically ordered by means of a subsumption relation, are also structured in terms of (disjunctive and non-disjunctive) opposition relations.

PAROLE-SIMPLE-CLIPS[7] (henceforth PSC[8]) is a four-layered lexicon developed over three different projects. Morphological and syntactic models and the kernel of related lexicons were elaborated in the EU *LE-PAROLE* project; the semantic model and the core of the semantic lexicon, in the EU *LE-SIMPLE* project[9]; the phonological level of description and the extension of the lexical coverage were performed in the context of the Italian national project *Corpora e Lessici dell'Italiano Parlato e Scritto* (CLIPS). The theoretical model underlying this lexicon is based on the EAGLES recommendations, on the results of the EWN and ACQUILEX projects and on a revised version of Pustejovsky's Generative Lexicon theory (Pustejovsky 1995).

At the semantic level, the PSC lexicon (Ruimy *et al.* 2003), which comprises more than 57,000 Italian word senses (verbs, common and proper nouns, adjectives, adverbs and grammatical words), is structured in terms of an ontology.

The SIMPLE Ontology[10] (SO) consists of 157 language-independent semantic types designed for the multilingual lexical encoding of concrete and abstract entities, properties and events. It is a multidimensional type system, based on hierarchical and non-hierarchical conceptual relations, which distinguishes between *simple* (one-dimensional) and *unified* (multi-dimensional) semantic types, the latter implementing the principle of *orthogonal inheritance* (Pustejovsky & Boguraev, 1993). Multidimensionality is captured by *qualia roles* that define the distinctive properties of semantic types and differentiate their internal semantic constituency.

Since IWN, unlike PSC, is a one-layer lexical database, the comparison of the resources focuses on their semantic information. In this regard, each lexicon provides a bundle of specific properties reflecting the different principles and peculiarities that characterize its underlying model[11] but also a large number of conceptually similar information that represent the compatible aspects of these two lexicons. In this connection, it is worth reminding that EWN was one of the inspiration sources for the SIMPLE model of semantic representation.

Studying the two resources, the wide range of compatibility observed did prompt us to undertake their semi-automatic link, eventually combining and merging the whole information into a common representation framework. In this respect, LMF, which enables the merging of electronic lexical resources, seems an appropriate candidate framework all the more since its creation was largely inspired by the PAROLE-SIMPLE model.

The remainder of this paper reports on the mapping of the ontologies on which basis both

[3] http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=834/vers=ing
[4] http://www.hum.uva.nl/~EWN
[5] The only aspects in which IWN differs from EWN are a few amendments made to the ontology in order to allow for the representation of adjectives and the addition of further lexical-semantic relations.
[6] Cf. note5.
[7] http://www.ilc.cnr.it/clips/CLIPS_ENGLISH.htm
[8] 'PSC' is not the acronym of the lexicon. It is only used here for brevity
[9] http://www.ub.es/gilcub/SIMPLE/simple.html
[10] http://www.ilc.cnr.it/clips/Ontology.htm
[11] such as, for example, a different ontological framework and a different approach to the organization of lexical units

lexicons are structured; it provides an overview of the linking methodology developed and of the mapping process implemented in an Access software tool; it illustrates the results of the first mapping phase, which was devoted to 1stOrderEntities. Resulting from the mapping, some reciprocal benefits for the resources are also illustrated that definitely justify the soundness of our linking initiative. Ongoing and future work is outlined in the conclusion.

3 Mapping the Ontologies

3.1 Ontological typing

Let us first very briefly illustrate the ontological typing in the two models.

According to the SIMPLE model, the basic unit, i.e. the word sense, represented by a 'semantic unit' (*SemU*) is associated to *one single* semantic type, e.g.: SemU69589*cardiologia* (cardiology) [DOMAIN]; USem4985*insegnante* (teacher): [PROFESSION]. Through its type membership, each *SemU* is endowed with a structured set of semantic features and relations; among these are the 60 relations of the *Extended Qualia structure*, a revisited version of the original GL representational tool that enables to describe both the componential aspect of a word meaning and its relationships to other lexical items.

The EWN/IWN model, by contrast, allows for a multi-classification. Synsets are in fact seldom linked to one single ontological node but rather cross-classified in terms of multiple, non-disjoint, TCs[12], e.g.: synset29146: {N *cardiologia*1}: [Agentive Purpose Social Undboundedevent]; synset4283: {N *docente*1, *didatta*1, *professore*2, *insegnante*1}: [Human Object Occupation]. Noteworthy here is that the ontological classification is determined by the choice of the synset hyperonym. As to the word sense or, in WN parlance, *synset variant,* its semantics is fully defined by its membership in a synset.

Although moving from a different approach to the definition of word sense, the information provided by these two types of ontological classification is substantially equivalent. Owing to the multidimensional nature of the ontology, SIMPLE types encompass in fact the various meaning dimensions that are expressed in IWN by the different TCs cross-classifying 1st and 2ndOrder Entities.

3.2 Mapping the ontological classes

In the process of mapping these two ontology-based lexical resources, the first step clearly consisted in comparing their ontological framework, viz. in manually establishing correspondences between the conceptual classes of both ontologies, with a view to further matching their respective instances. The comparison was done so far for classes structuring entities and events (Ruimy, 2005)[13]; the ontological typing of adjectives will be dealt with in a further phase of work.

A preliminary observation to be done is that IWN TO consists of a set of rather flat top semantic features whereas SO encompasses mono- and multi-dimensional types with associated templates of structured information that define the content of the conceptual types.

As mentioned in section 2, the first subdivision level of IWN TO consists of three main classes:

The 1stOrderEntity class structures concrete entities (referred to by concrete nouns). Its main cross-classifying subclasses: Form, Origin, Composition and Function correspond to the four Qualia roles the SIMPLE model avails of to express orthogonal aspects of word meaning. Their respective subdivisions consist of (mainly) disjoint classes, e.g. 'Natural' vs. 'Artifact', 'Substance' vs. 'Object'. To each class corresponds, in most of the cases, a SIMPLE semantic type or a type hierarchy subsumed by the CONCRETE_ENTITY top type. Some other TCs, such as 'Comestible' and 'Liquid', are instead mappable to SIMPLE distinctive features Plus_Edible, Plus_Liquid, etc.

The 2ndOrderEntity class classifies static or dynamic situations denoted by nouns and verbs, adjectives and adverbs. 2ndOrderEntities are primarily characterized in terms of two classification parameters: 'Situation Type' – whose two disjoint features, 'Static' and 'Dynamic' encode the event structure – and 'Situation Component', which subsumes a set of combinatorial classes providing a more conceptual classification in terms of semantic components of a

[12] The more specific the word, the more TCs contributing to its description.

[13] http://www.ilc.cnr.it/clips/Ontology_mapping.doc

concept, e.g.: 'Manner', 'Experience', 'Communication', 'Cause'.

Concerning the Situation Type, in the SIMPLE model, the event structure is expressed by means of the three-valued feature Eventtype = *state*, *process*, *transition*, values which correspond in IWN respectively to 'Static', (Dynamic) 'Unbounded-event' and (Dynamic) 'Boundedevent'. As to the combinatorial subclasses of the Situation Component, each one generally corresponds to one or more SIMPLE types, depending on the Situation Type value and/or the other Situation Components it combines with, as illustrated in table 1.

IWN Top Concepts	SIMPLE semantic type
Existence Bounded Cause Physical	CREATION, CAUSE_NATURAL_TRANSITION
Existence Static	EXIST
Experience Mental Dynamic	EXPERIENCE_EVENT, MODAL_EVENT
Experience Physical Stimulating Dynamic	PERCEPTION

Table 1. TCs combinations and semantic type correspondences

The 3rdOrderEntity class, which has no further subdivision, classifies abstract entities (denoted by abstract nouns) existing independently of time and space. These entities fall into the ABSTRACT_ENTITY type hierarchy of the SIMPLE ontology.

Notwithstanding the different approaches taken for their design and some different underlying principles, these two ontologies globally show a significant degree of overlapping and no fundamental difference in conceptualization is observed. Two general remarks are in order here:

1) Owing to the different extension of both ontologies, some specific concepts – which are expressed in SIMPLE Ontology by lower level semantic types – are likely to have no equivalent in IWN TO.

2) Not surprisingly, mapping from event-denoting PSC semantic units to IWN 2ndOrderEntities immediately appears more challenging than dealing with 1rstOrderEntities that pose less tricky problems.

4 Linking Methodology

Owing to a different organizational structure of information in the two resources, the linking process involves elements having a different status, viz. autonomous semantic units in PSC and synsets clustering 1 to n synset variants in IWN.

In order to avoid dealing with huge, unmanageable sets of data, mapping is performed on a semantic type-driven basis and is PSC → IWN oriented. The rationale for this orientation is that the 157 semantic types of the SO provide a more fine-grained structure of the lexicon than the 65 top concepts of the IWN ontology, which reflect only fundamental distinctions.

Taking therefore as starting point the lexical instances of a SIMPLE semantic type along with their PoS and hyperonymic ('isa') information, the IWN resource is explored in search of linking candidates.

Each mapping run returns two data sets:

► Matched pairs of word senses, i.e. SemUs and synset variants with identical string and PoS and whose ontological classification matches the correspondences established between the classes of both ontologies.

These word senses are linked after human validation.

Linking may occur between a *SemU* and a *one-variant* synset (1):

1. SemU66448bastone ↔ synset29146 {N *bastone*1}

or between a *SemU* and one word sense of a *multi-variant* synset (2):

2. SemU75412*adornare* ↔ synset35336 {V *adornare*1, *ornare*1, *decorare*1, *guarnire*3, *addobbare*4}

► Unmatched word senses, in spite of their identical string and PoS value. Matching failure may be due either to coverage discrepancies (lack, in IWN, of a lexical item or of the appropriate word sense corresponding to a PSC entry) or to a mismatch of ontological classification between word senses existing in both resources. Focusing on this latter case, two main obstacles hamper their matching:

1) An incomplete ontological information:

As already said, IWN synsets are cross-classified in terms of a combination of TCs. This combined notation is however sometimes only partially encoded and cases are not rare of 1rstOrderEntities lacking some meaning component or

2nd OrderEntities lacking one of their two classifying parameters.

For the linking purpose, the problem of incomplete ontological classification may, in a number of cases, be overcome by relaxing the mapping constraints. Yet, this solution can only be applied if the existing ontological classification, in spite of its incompleteness, is informative enough. More problematic to deal with are those cases of incomplete and little informative ontological labels. This is the case, for example, of 1rst OrderEntities as different as *medicinale, anello, laccio, vetrata* (medicine, ring, lace, glass window) and only classified as 'Function' or of 2nd OrderEntities lacking either a Situation Component or a SituationType, e.g. *unirsi* (to join) classified as 'BoundedEvent' or *sciogliere* (to melt) as 'Cause'.

2) A different ontological information:

Besides mere encoding errors for which a correction phase is foreseen, the ontological classification may be different with respect to the constraints imposed to the mapping run and the discrepancy may be imputable to:

i) A different but equally defensible meaning interpretation in each resource, e.g.: *ala* (aircraft wing): 'Part' vs. 'Artifact Instrument Object'. Word senses falling into this category are clustered into numerically significant sets according to their semantic typing and then studied with a view to establishing further equivalences between ontological classes or to identify, in their classification schemes, descriptive elements lending themselves to comparison.

ii) The presence of polysemous senses of the considered *SemUs* (e.g., USem65931*kiwi* 'Fruit' which is obviously discarded when mapping the *kiwi* instance of the 'Animal' class). Some of these word senses proceed from an extension of meaning, e.g. People-Human: *pigmeo, troglodita* (pygmy, troglodyte) or Animal-Human *verme, leone* (worm, lion) and are used with different levels of intentionality: either as a semantic surplus or as dead metaphors (Marinelli, 2006).

iii) A different level of specificity in the ontological classification, either due to the lexicographer's subjectivity or to an objective difference of granularity of the ontologies, cf. the *viola* example below.

Problems emerging with instances of iii) may be bypassed by climbing up the ontological hierarchy, identifying the parent nodes and allowing them to be taken into account in the mapping process.

Hyperonyms of matching candidates are also consulted during the mapping process and play a particularly determinant role in the resolution of cases whereby matching fails due to a conflict of ontological classification, namely:

- sets of word senses displaying a different ontological classification in each resource but sharing the same hyperonym, e.g. *collana, orecchino* (necklace, earring) are typed as CLOTHING in PSC and as 'Function' in IWN but share the hyperonym *gioiello* (jewel).
- polysemous senses belonging to different semantic types in PSC but sharing the same ontological classification in IWN, e.g.: in PSC, SemU1595*viola* (violet) PLANT and SemU1596*viola* FLOWER vs. in IWN: *viola*1 (has_hyperonym *pianta*1) and *viola*3 (has_ hyperonym *fiore*1) (flower), both typed as 'Group Plant'.

5 Mapping Process

The Access software tool devised to map the lexical units of both lexicons works in a semi-automatic way using the ontological classifications, the hyperonymic relations and some semantic features of the two resources. The mapping process foresees the following steps:

► Selection of a PSC semantic type and definition of the search range, i.e. either all of its instances or a subset bearing a selected feature, e.g. PLANT and 'Plus_Edible';

► Selection of one or more mapping constraints on the basis of the correspondences established between the conceptual classes of both ontologies;

► Human validation of the automatic mapping, i.e. selection of the semantically relevant word sense pair(s) from the set of possible matches automatically output for each *SemU* (referred to as *multiple mapping* in table 2). Multiple mappings depend on the more fine-grained sense distinctions performed in IWN. Cases are in fact frequent of a single entry in PSC corresponding to two different IWN entries encoding very fined-grained nuances of sense, e.g.: SemU63617*galeotto* vs. synset28576: {N *galeotto*1} (galley rower) and synset49579: {N *galeotto*2} (galley slave). The selection of the relevant word sense pair involves

checking information sources such as hyperonyms, SemU / synset glosses and ILI links;

► Relaxation / tuning or addition of mapping constraints, where appropriate; new processing of the input data.

6 Mapping Results

The results of the first working phase, which was devoted to linking concrete entities, sound quite encouraging since 72,32% of the word senses considered have been successfully linked. Table 2 evidences: i) the extent of overlapping coverage for concrete entities; ii) the considerable percentage of linked senses with respect to the linkable ones (i.e. words with identical string and PoS value); iii) the many cases of multiple mappings.

Overlapping coverage		56,29%
Selected *SemUs*	27,768	--
Linkable senses	15,193	54,71%
Linked senses	10,988	72,32%
Multiple mappings	1,125	10,23%
Unmatched senses	4,205	27,67%

Table 2. Mapping concrete entities

7 Enhancement of the Resources

Besides offering the end user a more exhaustive and in-depth lexical information combining the potentialities and most outstanding features of the two lexical models, the linking process lets inconsistencies that unavoidably exist in both resources emerge, allowing therefore to amend them. To give but an example, consistency would require that, when a *synset variant* is linked to a *SemU*, all the other variants from the same synset map to PSC entries sharing the same semantic type. Yet, especially concerning event denoting words, cases have been observed whereby *SemUs* corresponding to variants of the same synset do not share a common SIMPLE semantic type. Linking the two resources permits therefore to enhance their consistency since it implies a de facto reciprocal assessment of both coverage and accuracy, which is particularly relevant to hand-built lexical resources. 'Cleaning' the two lexical resources represents moreover a step forward towards their interoperability and eases therefore their eventual merging.

Moreover, the linking process makes it possible to enrich each resource by complementary information types that are peculiar to the other's theoretical model. In EWN, the richness of sense distinctions and the consistency of hierarchical links are remarkable. SIMPLE, on the other hand, focuses on richly describing the meaning and semantic context of a word and on linking its syntactic and semantic representation, which is crucial for most NLP applications.

7.1 IWN Information Enriching PSC SemUs

The organization of lexical knowledge has entailed a quite coherent clustering of synonyms in Wordnet-based resources. The SIMPLE model, on the other hand, has devoted more attention to other relation types and less importance has been given to the instantiation of synonymy links.

Integrating the two lexicons, PSC entries will easily be enriched by synonymy links, based on synset membership. Likewise, missing senses of existing words and new lemmas will be quickly and consistently encoded in PSC lexicon.

In IWN, hierarchical links are of fundamental importance and hence consistently expressed by two relations 'has_hyperonym' and 'has_hyponym' that allow a cross-checking of data. In the SIMPLE model, on the other hand, the focus put on covering the whole range of a word's syntactic and semantic uses has sometimes prejudiced the enforcement, in PSC entries, of coherent taxonomic links and yielded cases of circularity. Such cases will be amended by resorting to IWN hyperonymic links for nouns and verbs.

IWN 'Involved_agent / patient / instrument / location' and 'Role_agent / patient / instrument / location' semantic relations, respectively linking 2nd with 1rst or 3rdOrderEntities and conversely will be most helpful to relate more straightforwardly, in PSC lexicon, an event to its participants[14], e.g.: *operare*: involved_agent = *chirurgo*, involved_ patient = *paziente*, involved_location = *ospedale,* involved_instrument = *bisturi* (to operate, surgeon, patient, hospital, lancet); an entity to an event: *studente*: role_patient = *insegnare* (student, to teach) or even to relate event's participants to each other: *chirurgo*: co_agent_patient = *paziente.* These links will moreover allow to discriminate the nature of some relationships that are rather poorly rendered, in the PSC lexicon, by the overused — and hence misused — constitutive relation 'concerns', e.g.: 1)

[14] Work on this issue is now in progress (Ruimy, 2007).

otturazione concerns *dente* (filling, tooth)*;* 2) *sbarcare* concerns *nave* (disembark, boat) could be respectively replaced by involved_patient and **i**nvolved_source_direction.

7.2 PSC Information Enriching IWN Synsets

No argument structure information is provided in IWN. Linking the two lexicons, IWN predicative units will be endowed with information concerning their syntactic and semantic subcategorization frames.

IWN word senses will also inherit the PSC extensively encoded information concerning their domain of use. Such information, most relevant to IR, WSD, IE and parsing, enables – among other – clustering semantically related lexical items pertaining to specific domains, regardless of their PoS and type membership.

Given the rich lexical information foreseen by the SIMPLE model, IWN synsets will also gain:

- a finer-grained ontological classification: let us observe for example that, as against the 'Plant', 'Human', and 'Communication' TCs, SIMPLE Ontology offers respectively 5, 9 and 8 semantic types, each one providing a rich bundle of specific information;
- a semantic description less prominently based on taxonomic relations. SIMPLE semantic relations, which are defined along multiple dimensions, enable to avoid an overloading of the 'isa' relation and to represent senses not adequately definable in terms of the hyperonymic link.
- the expression of further orthogonal meaning dimensions: e.g., synset variants such as *inchiostro* or *colla* (ink, glue), associated to the TC 'Substance' and bearing telic information, will acquire, through their linking to the corresponding PSC entries, constitutive and agentive dimensions.
- the account of systematic polysemy. Regular polysemy is expressed, in the SIMPLE model, through distinct entries connected by means of a polysemous relation linking the ontological typing of pairs of senses, according to a set of polysemous classes, e.g. *banca* (bank): BUILDING / INSTITUTION; LOCATION / HUMAN_GROUP. In IWN, such polysemous senses are encoded as separate meanings but no mention is made about the way they relate to each other. The possibility provided by the EWN model to assign two or three disjunctive hyperonyms was in fact not systematically implemented. During the IWN project, a proposal was made to encode regular polysemy using the 'is_extension_of / has_extension' relation, originally created for proper nouns (Marinelli, 2004).
- a more specific identification of the nature of some syntagmatic relationships not expressible in the IWN model, and which are, for instance, most relevant for extracting semantic networks, e.g.: *antidoto* 'used_against' *veleno* (antidote, poison), *acetone* 'used_as' *solvente* (acetone, solvent).

8 Conclusion and Future Work

This paper reported on the ongoing linking of the two largest general-purpose, electronic lexical resources of Italian language, PAROLE-SIMPLE-CLIPS and ItalWordNet. The mapping of the ontologies on which basis the lexicons are structured was described. An overview of the adopted methodology, of the linking process and of the results of the first mapping phase regarding 1stOrderEntities was provided. Reciprocal benefits and enhancements for the two resources were also illustrated.

Differences regarding the nature of linking units, the granularity of sense distinction and the ontological typing are complex issues that are being addressed during the linking process. Problems arise, in particular, when encoding incompleteness or inconsistency generate unpredictable, non-systematic ontological typing discrepancies whereby a theoretical comparison of the models evidenced a high degree of overlapping. Nevertheless, the wide range of compatibility the models show induces us to believe that semantic interoperability is indeed achievable and it is our strong conviction that linking two resources based on such valuable and widely tested lexical models that have addressed challenging (and complementary) research issues in lexical semantics is a most appropriate and timely initiative. Semantic integration of these resources is all the more desirable considering their multilingual vocation: IWN is linked to the WN of seven other languages and PSC shares with eleven European lexicons a theoretical model, representation language, building methodology and a core of entries.

On the basis of the encouraging results obtained from the linking of 1rstOrderEntities, work is now in progress as regards the mapping of

3rdOrderEntities and 2ndOrderEntities which, so far, had only been object of preliminary investigations on Speech act (Roventini, 2006) and Feeling verbs.

References

Adriana Roventini, Marisa Ulivieri and Nicoletta Calzolari. 2002. *Integrating two semantic lexicons, SIMPLE and ItalWordNet: what can we gain?* LREC 2002, Third International Conference on Language Resources and Evaluation Proceedings, Vol. V, pp. 1473-1477, Las Palmas de Gran Canaria.

Adriana Roventini et al. 2003. ItalWordNet: *Building a Large Semantic Database for the Automatic Treatment of Italian*. Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI, vol. II, 745-791.

Adriana Roventini and Nilda Ruimy 2005. *Linking and harmonizing different lexical resources: a comparison of verbal entries in ItalWordNet and PAROLE-SIMPLE-CLIPS.* In Proceedings of the Third International WordNet Conference, Jeju Island, Korea

Adriana Roventini. 2006. *Linking Verbal Entries of Different Lexical Resources*. LREC Proceedings, CD-ROM, 1710-1715.

Alessandro Lenci et al. 2000. *SIMPLE Linguistic Specifications*, Deliverable D2.1, ILC-CNR, Pisa.

Antonietta Alonge, et al. 1998. *The Linguistic Design of the EuroWordNet Database*, Special Issue on EuroWordNet. In N. Ide, D. Greenstein, P. Vossen (eds.), 'Computers and the Humanities', XXXII, 2-3, (91--115).

Christiane Fellbaum (ed.) 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

George Miller et al. 1990. *Introduction to WordNet: An On-line Lexical Database*, International Journal of Lexicography, III, 4, 235--244.

Gil Francopoulo, et al. 2007. *Lexical Markup Framework: an ISO Standard for Semantic Information in NLP Lexicons*, Proceedings of the Workshop on Lexical-Semantic and Ontological Resources of the GLDV Working Group on Lexicography at the Biennal Spring Conference at the GLDV, Tübingen, (13-14/04/2007).

James Pustejovsky and Branimir Boguraev 1993. *Lexical knowledge representation and natural language processing,* Artificial Intelligence, Volume 63, Special volume on natural language processing, Issue 1-2, 193 – 223.

James Pustejovsky 1995. *The generative lexicon*. The MIT Press, Cambridge, MA.

James Pustejovsky 1998. *Specification of a Top Concept Lattice*, ms., Brandeis University.

Nilda Ruimy et al. 2002. CLIPS, *a Multi-level Italian Computational Lexicon*, LREC 2002, Third International Conference on Language Resources and Evaluation Proceedings, Vol. III, (pp.792-799), Las Palmas de Gran Canaria.

Nilda Ruimy et al. 2003. *A computational semantic lexicon of Italian: SIMPLE*. In A. Zampolli, N. Calzolari, L. Cignoni, (eds.), Computational Linguistics in Pisa, Special Issue, XVIII-XIX, (2003). Pisa-Roma, IEPI. Tomo II, 821-864.

Nilda Ruimy and Adriana Roventini. 2005. *Towards the linking of two electronic lexical databases of Italian*, In Z. Vetulani (ed.), L&T'05, April 21-23, Poznan, Poland. Wydawnictwo Poznanskie Sp. z o.o. 230-234.

Nilda Ruimy 2006. *Merging two Ontology-based Lexical Resources*. LREC Proceedings, CD-ROM, 1716-1721.

Nilda Ruimy 2007. *Enhancing SIMPLE Semantic Relations: A Proposal***.** In Zygmunt Vetulani (ed.), Proceedings of 3rd Language & Technology Conference. Fundacja Uniwersytetu im A. Mickiewicza, Poznań. 119-123.

Nuno Silva et al. 2005. *An approach to ontology mapping negotiation*, Proceedings of the Third International Conference on Knowledge Capture Workshop on Integrating Ontologies; Banff, Canada.

Piek Vossen (ed.) 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers.

Piek Vossen (ed.) 2002. *EuroWordNet: General Documentation*, Version 3, Final, July 1.

Rita Marinelli. 2004. *Proper Names and Polysemy: From a Lexicographic Experience*. LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, Proceedings, Volume I, 157-160.

Rita Marinelli. 2006. *Computational Resources and Electronic Corpora in Metaphors Evaluation*. Second International Conference of the German Cognitive Linguistics Association, Munich, 5-7 October.

SHACHI: A Large Scale Metadata Database of Language Resources

Hitomi Tohyama[†], Shunsuke Kozawa[†], Kiyotaka Uchimoto[††], Shigeki Matsubara[†] and Hitoshi Isahara[††]

†Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

{hitomi,kozawa,matubara}@el.itc.nagoya-u.ac.jp,

††National Institute of Information and Communications Technology

3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan

{uchimoto,isahara}@nict.go.jp

Abstract

The National Institute of Information and Communications Technology (NICT) and Nagoya University have been jointly constructing a large scale database named SHACHI by collecting detailed meta information on language resources in Asia and Western countries, for the purpose of effectively combining language resources. The purpose of this project is to investigate languages, tag sets, and formats compiled in language resources throughout the world, to systematically store language resource metadata, to create a search function for this information, and to ultimately utilize all this for a more efficient development of language resources. This metadata database contains more than 1,700 compiled language resources such as corpora, dictionaries, thesauruses and lexicons, forming a large scale metadata of language resources archive. Its metadata, an extended version of OLACmetadataSet conforming to Dublin Core, which contain detailed meta information, have been collected semi-automatically. This paper explains the design and the structure of the metadata database, as well as the realization of the catalogue search tool.

1 Introduction

The importance of the construction of language resources such as corpora, dictionaries, thesauruses, etc., has been widely recognized, and has boomed for years throughout the world in its aim of encouraging research and development in the main media of spoken and written languages. Among the organizations willing to store and distribute language resources, there exist some consortia fulfilling their function such as LDC (Linguistic Data Consortium), ELRA (European Language Resources Association), OLAC (Open Language Archives Community) and Chinese-LDC (Chinese Linguistic Data Consortium) in Western countries and China, and GSK (Gengo Shigen Kyokai; Language Resource Association) which does so mainly in Japan. However, those released language resources are scarcely connected with each other not only because of the difference between written languages and spoken languages but also because of the difference between languages such as Japanese, English, Chinese, etc. Moreover, since language information tags given to those language resources and their data formats are multifarious, each language resource is operated individually.

As language resource development generally entails enormous cost, it is highly desirable that the research efficiency be enhanced by combining those existing language resources and systematically developing them altogether. For the purpose of fully integrating their language resources, NICT and Nagoya University have been constructing a large scale database named SHACHI as their joint project by collecting detailed meta information on language resources in Western and Asian countries.

This research project aims to extensively collect, and systematically store, metadata such as tag sets, formats, and language resource recordings existing at home and abroad. Meanwhile, we have already

developed a facet search function of language resources using meta information, and are performing the experiment of widely providing this meta information on the stored language resources to those from professional researchers to common users. This paper outlines our language resource database, named SHACHI, in the development stage.

The structure of this paper is as follows: first we will outline the purpose and design of SHACHI in the second chapter. Next we will describe the collection of metadata in the third chapter, and the database structure, as well as the fundamental statistics, in the fourth chapter. Finally we will explain our future task in the fifth chapter.

2 Design of the Metadata Database

2.1 Purpose of the Construction of SHACHI

The purpose of the construction of SHACHI is fourfold.

- To investigate the actual conditions of tags and format types of language resources existing at home and abroad.
- To systematically obtain and store metadata of international language resources according to the information obtained from the above-mentioned. (This leads to the construction of language resource ontology.) (Hayashi, 2007)
- To conduct an investigation into the organic combination of language resources. (This leads to the strategic development of language resources.)
- To promote the distribution of language resources.

Some 2,000 resources of metadata have already been collected in the database so far, and by December 2008 they will be further enlarged by 3,000. To that end, it is indispensable for us to work in cooperation with language resource consortia at home and abroad, and to take the initiative in contributing to Asian language resources.

Additionally, this database is obviously different from those of other language resource consortia since all of our detailed metadata are inputted manually. The database is notably characterized by the attempt to make an ontological construction of language resources throughout the world, as the affinity of language resource types and that of their tag sets are analyzed by applying natural language processing techniques to those detailed metadata. It seems certain that the realization of its ontological construction will contribute to a cutback in research and development costs, and to establishing a language resource infrastructure which meets present-day needs as an on-demand service. (NICT, 2007).

2.2 Design for Collecting Metadata

Among organizations willing to store and distribute language resources, there exist some consortia fulfilling their function such as LDC, ELRA, OLAC and Chinese-LDC in Western countries and China, and mainly GSK in Japan.

As for websites, there are two attempting to systematically amass metadata of language resources and promote their distribution, such as Language Technology World (LTW 2007) owned by DFKI (Deutsches Forschungszentrum for Kunstiliche Intelligencz) and a page owned by OLAC (OLAC, 2007).

To return the benefit of developed information processing technologies to society, it is highly desirable that the research be done in mutual cooperation among various language resource consortia and be enhanced by mutually exchanging information. SHACHI will make this possible as its metadata enables us to collect more detailed meta information in accordance with the OLAC metadata set by extending it. OLAC is creating a worldwide virtual library of language resources by developing consensus on the best current practice for the digital archiving of language resources, and by developing a network of interoperating repositories and services. OLAC metadata is based on the complete set of Dublin Core metadata set but a part of which was extended. The metadata set of SHACHI is described in detail in the following third chapter.

Table 1. SHACHI metadata set

	DC Qualifiers		Qualifiers for more precise description of the resources	
	LEVEL 1	LEVEL 2		
	DCMES Elements	DC Element Refinements	OLAC Extensions	SHACHI Extensions
1	title	alternative		
2	creator			
3	subject		linguistic-field (29) language (OLAC-Language extension)	mono-multi-lingual (2) monolingual multilingual resource-subject (3) dictionary thesaurus lexicon
4	description			price
5	publisher			
6	contributor		role(24)	mother-tongue (2) native non-native intonation (2) standard_dialect dialect level (2) age (3) gender (3)
7	date	created issued		
8	type		discourse-type (10) linguistic-type (3)	purpose(4) lexicography analysis developing_technologies education style (2) speech written form (2) fixed unfixed sentence(3) short long mixed has-annotation (2) annotated plain annotation sample
9	format	extent medium		encoding markup functionality
10	identifier			
11	source			
12	language		language (OLAC-Language extension)	
13	relation	DC relation refinements (13)		utilization
14	coverage	temporal		
15	rights			

Table 2. Description of SHACHI extensions

SHACHI Extensions	Description
mono multi lingual (2)	Applies to: subject
monolingual	A resource using only one language:. The same language is use for the subject language and to describe the subject language.
multilingual	A resource in several different languages. Different language(s) are used for more than one subject languages and to describe the subject language(s).
ResourceSubject (3)	Applies to: subject
dictionary	A list of the words of a language in which the definitions or meanings of the words are explained either in the same language or in a different language.
thesaurus	A list of the words of a language in which the words are arranged in groups that have similar meanings.
lexicon	A list of words on a particular subject.
Attribute (5)	Attributed of a contributor
mother-tongue	The performer is whether a native or non-native speaker of the language.
intonation	Dialectal status whether the performer uses a standard language or a dialect.
level	Whether the performer has received a professional level of linguistic(speaking or writing) training or no such training.
age	An age group the performer is in. When there are many performers in a resource, the ratio between all the age groups to which the performers
gender	Sex of the performer. When there are many performers in the resource, the ratio of males and females.
Purpose(4)	Applies to: type
lexicography	The creation of the resource is intended for lexicography.
analysis	Thecreation of the resource is intended for analysis.
developing technologies	Thecreation of the resource is intended for developing technologies.
education	The creation of the resource is intended for the use in education.
Style (2)	
speech	The resource is of the spoken language.
written	The resource is of the written language.
Form (2)	
fixed	A collection of fixed forms of expressions.
unfixed	The resource collects various forms of expressions.
Sentence(3)	
short	A collection of short sentenses.
long	A collection of long sentenses.
mixed	A collection of varying length of sentences.
Annotation (3)	
annotated	Tagged corpus.
plain	A corpus without annotations.
annotaion_sample	A sample of tagged data
Sample	A sample of the language resource
Format (2)	
encoding	An encoded character set used by a digital resource.
markup	A markup scheme used by a digital resource.
Functionality	Software Functionality
Utilization	Applicability of the resource. The described resource is utilized for the referenced technology, education, research or a product.

3 Collecting Metadata

3.1 Extensions of Metadata Element

The metadata set of this language resource database follows 15 kinds of elements of Dublin Core (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, rights), which is an extended version of OLAC metadata set, Table 1 shows the SHACHI metadata set. The shaded boxes "SHACHI Extension" are the elements added to those of Dublin Core. As for the definition of extended metadata elements, Table 2 is provided for showing it with the proviso that only one of the extended elements named "Utilization" is adopted on trial. This element is to provide information on the use of language resources which is semi-automatically retrieved from scholarly papers (Kozawa, 2007). Those items were added to this metadata, because it seemed that they gave us clues for finding the attributes of a corpus in collecting metadata of language resources such as corpora. Furthermore, we encourage our registrants to register as minutely as possible the information corresponding to each metadata element by them-

selves instead of automatically. In consequence, the information sometimes contains a lot of key words and somewhat longer sentences, which however enlarges the data subject to the key word search. We believe that its information will provide important clues to measure the affinity (similarity) of each language resource through language processing technologies in the future. The development of this project must contribute to the construction of an international language resource ontology.

As for the meta elements of language attributes, 160 values are currently introduced in our metadata set, which conforms to ISO 639 (ISO, 2007). The way to describe the date and the time conforms to ISO 8601. Moreover, this database will be intended to conform to ISO TC46/SC4, the International Standard for Information and Documentation & Technical Interoperability.

3.2 Policy for Collecting Metadata

The language resources which SHACHI stores should satisfy the following conditions:

1. The resources should be stored in a digital format.
2. The resources should be one of the followings: corpus, dictionary, thesaurus, or lexicon. (Numeric data are not considered to be the subject of collection for this database.)
3. Those resources should be collected from English websites and its data must be open to the public.
4. Those resources should be created by research institutions, researchers, or business organizations.

In addition to the above conditions, we are primarily collecting metadata of language resources which contain a large volume of data, are well known to the public, and are considered to be important for improving information processing technology. As expected we will call for participation in our project from research laboratories and institutes developing languages resources, by registering them in our database, as soon as preparations for publishing RSS feeds of our data and our input format of metadata standards are completed. Its registration page is also to be open to the public but is not yet on view.

- **Preferred Collection of Highly Recognized Language**

  It is essential to obtain and store information on highly recognized and frequently used language resources through collecting language resource metadata. Therefore, during the construction of SHACHI, a web search through Google was first conducted to investigate the actual state of language resources that meet the above mentioned conditions. Then, a search for language resources using key words such as “corpus”, “dictionary”, “thesaurus”, “translation” or “multilingual”, or key phrases such as “corpus so-and-so” (e.g. such as ‘PennTreeBank’), was done and as a result those which ranked highly and fitted our key words were retrieved. Table 3 shows the result. According to this research, WordNet was found to be the most widely distributed language resource used as a thesaurus. WordNet was originally compiled in certain European languages, and is now also being compiled in Chinese, Korean and other Asian languages. On the other hand, it was found that WordNet in Japanese has not been on the Internet yet. Research done through the Internet such as that described here is considered to be important in identifying the types of language resources to be developed in the future.

- **Collecting the Metadata from Major Organizations**

  It is important to gather language resource metadata from all over the world to perform a study about language resources, their promotion of distribution, and their strategic development. SHACHI not only covers metadata from major language resource consortia in Japan, Western countries, and China, but also conducts semi-automatic ways of registering detailed metadata in accordance with SHACHI metadata sets. Table 4 shows the list of major language resource consortia which this database covers.

4 Construction of SHACHI

SHACH is composed of language resource catalogs, a list of all language resource catalogs, catalogue search tool by which users can retrieve the information from all language resources stored in SHACHI from all angles, and the statistical information of the metadata of the language resources of SHACHI.

Table 3. The results of investigation on highly recognized and frequently used language resources

Monolingual Dictionary	Multilingual Dictionary	Parallel Corpus	Thesaurus
Webster's Revised Unabridged Dictionary (1913)	The EDICT Dictionary File	Aligned Hansards corpus	WordNet
Oxford Advanced Learner's Dictionary	WebLSD	European Parliament Proceedings Parallel Corpus 1996-2003	EuroWordNet
LONGMAN Dictionary of Contemporary English	Oxford-Hachette French Dictionary	OPUS - an open source parallel corpus	Hindi WordNet
Wiktionary	Collins ROBERT French Dictionary	UN Parallel Text	Roget's
Collins COBUILD Advanced Learner's Dictionary	Equine Multilingual Dictionary	Hong Kong Parallel Text	MeSH
The Swedish PAROLE Lexicon	The Papillon Project	COMPARA	Global WordNet
American Heritage Dictionary of the English Language	CJK Lexical Resources	The English-Norwegian Parallel Corpus	UMLS Metathesaurus
EDR Electronic Dictionary Technical Guide	Multilingual Dictionary of Proper Nouns CJKE-DPN	CRATER Multilingual Aligned Annotated Corpus	Merriam-Webster's Collegiate Thesaurus
Le Petit Robert French monolingual dictionary	Multilingual Glossary of technical and popular medical terms in nine European Languages	The JRC-Acquis Multilingual Parallel Corpus	ERIC Thesaurus
SANSEIDO Daijirin	*EIGIRO*	Polyglot Bible	Art & Architecture Thesuarus
A Japanese Lexicon (Japanese link)	EDR Bilingual Dictionary	NICT JLE Corpus	The European multilingual thesaurus on health promotion in 12 languages

Table 4. List of major language resource consortia which SHACHI covers

Consortium	ULR
Asian Language Resource Catalogue	http://nlp.kuee.kyoto-u.ac.jp/
ChineseLDC (ChineseLinguistic Data Consortium)	http://www.chineseldc.org/
ELRA(European Language Resources Association)	http://www.elra.info/fr/
Global WordNet Association	http://www.globalwordnet.org/
GSK (Gengo Shigen Kyokai)	http://www.gsk.or.jp/
ICAME Corpus Collection on CD-ROM	http://icame.uib.no/newcd.htm
LDC(Linguistic Data Consortium)	http://www.ldc.upenn.edu/Catalog/
Natural Language Processing Portal Site	http://nlp.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-j.html
SiTEC (Speech Information Technology & Industry	http://www.sitec.or.kr/English/
Speech Resources Consortium	http://research.nii.ac.jp/src/links/

The input format of metadata is, of course, not open to the public, but will be released in the future for use by research organizations and language resource consortia scattered all over the world. Because of this, we intend to provide outfits and interfaces which will allow researchers to register their language resource metadata without restraint, and which will enrich the quality of SHACHI by extending the contents of its database.

4.1 Outline of Catalogue

This catalogue allows SHACHI users to obtain outlines of the language resources at a glance without referring to the linked official websites (See Figure 1).

SHACHI catalogue allows SHACHI users to grasp outlines of the language resources at a glance without referring to the linked official websites.

4.2 Catalogue Search Tool

A catalogue search tool has been developed for the purpose of allowing users who visit the SHACHI site to find a language resource that meets their needs. This search tool is loaded with a key word search function as well as a facet search function. Figure 2 shows an image screen of the search tool. As for the key word search function, it allows users to input key words as they want, and to search all words stored in the SHACHI metadata archive. On the other hand, the facet search function is equipped with choices of 15 kinds of metadata elements selected from 25 kinds that SHACHI stores. This function helps users to obtain the desired language resource by selecting, in order, an element which is closest to their needs and then allows them to narrow down their choices. SHACHI especially brings its ability into full play for users who have no idea of adequate key words or have only a vague idea of how to find their desired information, since it can provide a large scale database with more detailed metadata through its characteristic search function (Bontcheva, 2007). On an actual image screen of the search results, titles of language resource candidates and a list of their descriptions are displayed, and at the bottom the relationships among other language resources are indicated (See Figure 2). We intend to conceive a method which enables relationships among language resource candidates

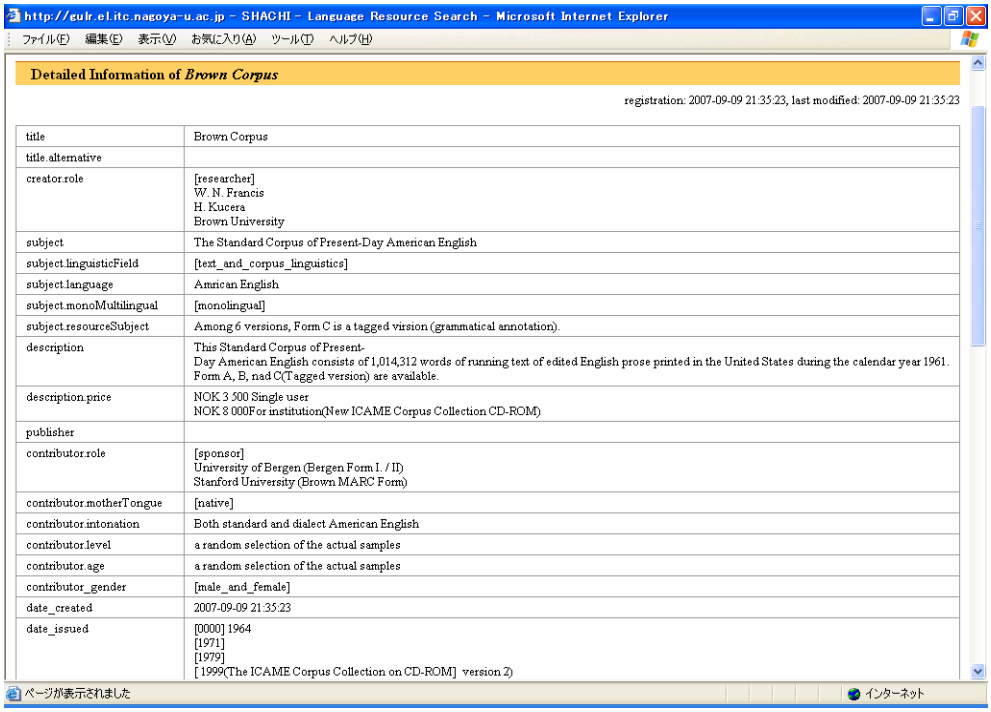


Figure 1. A sample page of SHACHI catalog (A part of catalogue; ex Brown Corpus)

shown in the search result to be visualized and measured.

4.3 Statistical Data

By observing metadata of the collected language resources, it is possible to find the transition of characteristics possessed by language resources that have been released so far. This website provides on the spot statistical information derived from the language resource metadata database SHACHI. It gives us the information, for example, that Arabian, Korean and Chinese language resources have been increasing recently, and that the prices of language resources have been rising. Thus, the SHACHI project investigates what kinds of language resources are needed and how they can be efficiently developed.

5 Future Work

These days new language resources are being created one after the other, and those resources are also modified on a daily basis. In order to grasp the current conditions of those resources, collecting and updating metadata at regular intervals is inevitably required. In addition, it is indispensable for us to collaborate with other research institutions such as GSK in order to collect language resource metadata efficiently and to develop the half-automated way of collecting and updating information via Web conversation (an online communication tool). On the other hand, we will conceive of an automatic information retrieval method by which we will obtain information on the situation of language resource diffusion, and various ways to utilize this information collected through the results of web searches and quoted information from papers.

6. Conclusion

The National Institute of Information and Communications Technology and Nagoya University have been collecting meta information extensively in order to construct a new type of language resource which uses metadata in their research through an organic combination of language resources. In this paper, the design of SHACHI (a metadata database of language resources now being developed), the expansion and construction of metadata for it, and the actualization of a search function we have developed were reported.

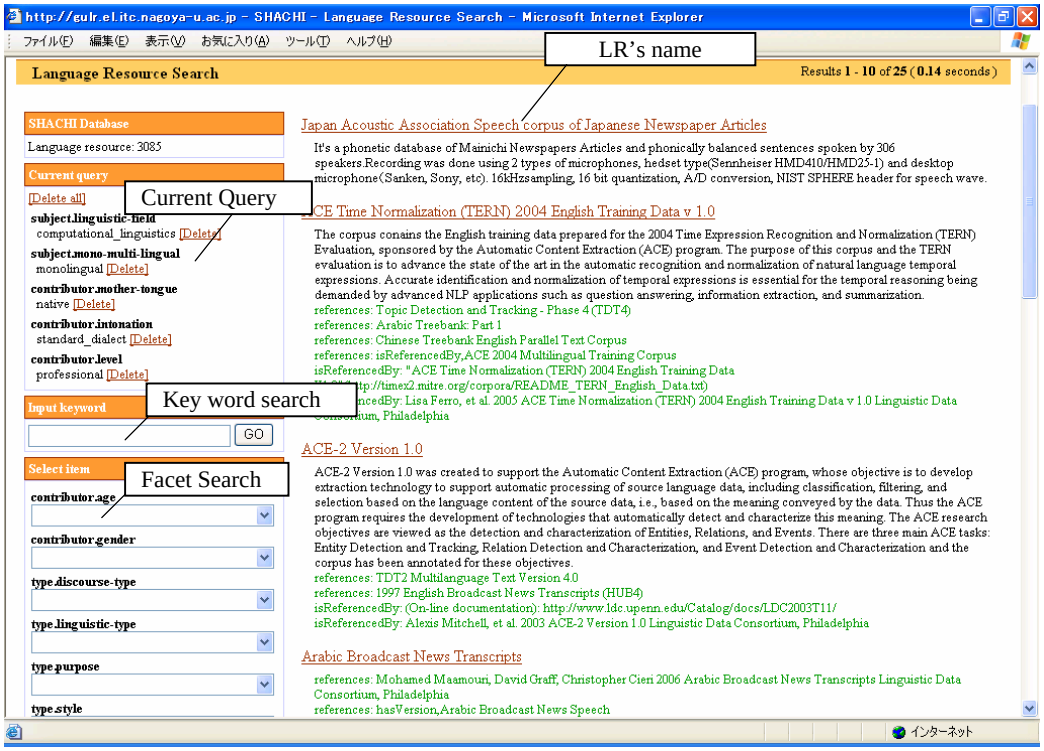


Figure 2. Catalogue search tool

SHACHI contains extensive data obtained from highly recognized language resources searched for on the Internet, and also covers meta information provided by language resources consortia such as ELRA, OLAC and LDC. Furthermore, this continuing work to register more detailed metadata is making SHACHI an even larger scaled database. At present, it contains approximately 1,700 pieces of meta information on language resources, forming the world largest language resource metadata archive. One of SHACHI's characteristic features is that the manner in which it collects tag sets and format samples given to language resources, has a desirable design for the strategic development of language resources including the standardization of tags and the efficient development of language resources. This collection of meta information has been further reinforced by the manual entering of even more detailed meta information into the bottom-level category of the meta element set. We have already measured the affinity between language resources and have systematically stored and classified language resources throughout the world. We believe this will lead to the construction of a language resource ontology.

Acknowledgements

We would like to thank Sachiko Waki, Miho Ohnishi of Nagoya University, Kenji Sugiki and Takahiro Ono of Matsubara's SLP Group at Nagoya University and all the staff of ANCHOR Co., Inc. for their full cooperation in constructing the database.

References

Hayashi, Y. 2007. Conceptual Framework of an Upper Ontology for Describing Linguistic Services, Proceedings of IWIC2007, pp.246-260.

ISO official site, 2007. http://www.iso.org/error/sitedown.html

Kozawa, S. Toyama, H. Uchimoto, K. Matsubara, S. 2007. Automatic Acquisition of Expressions Representing Purposes and Methods of Using A Language Resource from Academic Articles, Proceedings of Workshop on Informatics 2007, pp.65-70, (in Japanese).

LTW (Language Technology World). 2007. official site, http://www.dfki.de/lt//publications_show.php?id=148

NICT Language Grid project official site, 2007 . official site, http://langrid.nict.go.jp/indexj.htm.

OLAC (Open Language Archives Community), 2007. official site, http://www.lt-world.org/

EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data

Andrejs Vasiljevs
Tilde / University of Latvia
Vienibas gatve 75a
Riga, LV1004, Latvia
andrejs@tilde.lv

Signe Rirdance
Tilde
Vienibas gatve 75a
Riga, LV1004, Latvia
signe.rirdance@tilde.lv

Andris Liedskalnins
Tilde
Vienibas gatve 75a
Riga, LV1004, Latvia
andris.liedskalnins@tilde.lv

Abstract

This paper presents approaches in consolidation of dispersed multilingual terminology resources based on experience of EuroTermBank project. EuroTermBank is a new type of standards-based multilingual terminology gateway that provides unified access to centralized and distributed multilingual terminology resources, applying existing and emerging standards and concepts to ensure interoperability on several levels across diverse terminology data. Its federation approach enables consolidation of centrally stored resources with resources residing in diverse established terminology databases. The concept of terminology entry compounding provides for consolidated representation of entries across multiple terminology collections. Broader context of EuroTermBank activities is provided in the perspective of terminology consolidation at local, national and international levels of terminology work. TeDIF and TBX standards and their application in EuroTermBank are described.

1 Introduction

The requirement towards greater interoperability of terminology resources is determined by technological advances in language technologies and machine translation, as well as by advances in globalization of economy and growing global communication. The sound basis for consolidation, harmonization and general availability of international terminology is reliance on existing standards and, where existing standards are insufficient, their improvement and development of new standards. The multiple application areas of terminology resources demand their interoperability on several levels, starting from metadata description of terminology collections to common interchange formats.

This paper looks at the results of the EuroTermBank project (Auksoriute, 2006) in consolidation of dispersed multilingual terminology resources and analyzes how greater interoperability has been achieved through application of existing standards and new approaches.

2 Terminology consolidation on local, national and international levels

Terminology work is organized differently in different countries, institutions and settings. These differences are determined by individual goals and objectives of the involved organizations. Generally, we can distinguish 3 levels or scenarios of terminology work – local, national and international (Henriksen et al. 2005). In this section we will briefly characterize these levels in the aspect of terminology consolidation and interoperability based on assessment carried out in the EuroTermBank project.

2.1 Organizational level

On the local level, terminology work serves the needs of a particular organization, such as a company, a translation agency, a documentation centre, a research institute, etc. The local level is usually limited to one or a few closely related domains and is primarily concerned with terminology work originating from translation or creation of documents. Terminology work at the local level is usually limited in scope and the involved people, therefore terminology consolidation is not a major concern at this level (exceptions are multi-national companies and institutions with terminology work spread

around the globe). Although interoperability is not among the top priorities at this level, there is a growing awareness about the potential benefits from integration of locally created terminology into common terminology infrastructure.

2.2 National level

This level is concerned with the monolingual or bilingual terminology work performed on the level of a specific country. Among basic tasks of institutions involved in national terminology are national standardization and approval of terms, maintenance of national terminology, and development of integrated terminology systems

In some countries like Latvia and Lithuania, national terminology work also serves the normative purpose defining the "official" terminology for use in legislative documents. In other countries, consolidation and coordination are the major foci at the national level.

Exchangeability and harmonization of terminology resources typically are of high to medium priority at this level, as it may involve a complex structure of actors, compliance to national regulatory management of a national term database and a harmonized multi-branch term system.

Let us mention some examples of terminology consolidation at the national level. The Latvian terminology database[1] termnet.lv is maintained by Terminology Commission of Academy of Sciences of Latvia. It provides access to about 145 000 terms of different domains. It is also a portal where all new official terms get posted, and users can comment them.

In Lithuania, the State Language Commission together with the Chancellery of the Parliament took the initiative to create the State Bank of Terms; the law on the Term Bank of the Republic of Lithuania was passed in 2003. The purpose of the Term Bank[2] is to ensure a consistent usage of normalized Lithuanian terms, especially in the legislative documents of the Republic of Lithuania, to create a common informational system for various state institutions, including the option for other persons and legal entities to get connected to it and provide data to it.

2.3 International level

The international level concerns consolidation and harmonization of terms coined at the national and local levels; it involves coordination and management of multilingual terminology in a well-organized infrastructure. Since consolidation of terminology resources is the cornerstone of terminology work at this level, it not only requires rigorous application of existing standards, but also acts as the driving factor behind improvements and development of new standards and approaches.

Terminology collections at the international level are multilingual, this being a differentiator from other levels that are usually focused on one or a few languages. Terminology work at the international level should optimally include coordination of terminology work between the different countries and institutions involved as well as ensure data interoperability and facilitate terminology harmonization.

A good example of terminology consolidation and harmonization on a supranational scale is IATE[3] (Inter-Active Terminology for Europe), the EU inter-institutional terminology database used internally by the EU institutions and agencies since summer 2004 for the collection, dissemination and shared management of EU-specific terminology, and released for public access in 2007. It incorporates all of the existing terminology databases of the EU's translation services into a single new, highly interactive and accessible inter-institutional database. Its goal is to provide a centralized system for all EU terminology resources as a single access point that serves the EU institutions as well as EU citizens.

Important steps towards an interoperable model of terminology management within an international organization are taking place in ISO. The ongoing project of developing the ISO Concept database envisions a federated approach to development and maintenance of content, as well as public access to ISO terminology, in the form of ISO electronic dictionary (Weissinger, 2007).

2.4 EuroTermBank general approach

EuroTermBank project[4] is targeted to facilitate terminology data accessibility and exchange at all

[1] http://www. termnet.lv
[2] http://terminai.vlkk.lt
[3] http://iate.europa.eu
[4] http://www.eurotermbank.com

three levels. The initial focus of the EuroTermBank was to contribute to improvement of the terminology infrastructure in the selected new European Union member countries (Latvia, Lithuania, Estonia, Poland, Hungaria) but project expands its activities to other countries in EU and beyond. This aim is accomplished by establishing cooperative networks of terminology institutions on various levels and by consolidation and harmonization of existing terminology resources resulting in multilingual online terminology base.

EuroTermBank enables the exchange of terminology data with existing national and EU terminology databases by establishing cooperative relationships, aligning methodologies and standards, designing and implementing data exchange mechanisms and procedures. Through harmonization, collection and dissemination of public terminology resources, EuroTermBank is aimed to facilitate enhancement of public sector information and strengthen the linguistic infrastructure in the new EU member countries.

Development, population and maintenance of a web-based terminology data bank constitute the major tangible outcome of the project. The data bank works on a two-tier principle – as a central database and as an interlink node or a gateway to other national and international terminology banks.

The objective of EuroTermBank is to serve as a platform for consolidation of dispersed terminology resources into the central EuroTermBank database or interlink them via EuroTermBank as a central gateway and a single point of service.

3 Standards based data modeling for interoperability

With the multitude of actors involved in terminology work, implementation of applicable international standards, as described in this section, is a key to reaching the goals of the project, including a standards-based approach to describing terminology collections, defining the data model and ensuring a unified data exchange format.

3.1 Unified format for annotation of terminology resources

One of the major tasks in terminology data consolidation is identification and description of terminology resources. Due to a large number of resources to be described and different organizations in several countries involved it is important to use a common format for resource description. For this purpose we propose to use TeDIF format (Betz, Schmitz, 1999).

The Terminology Documentation Interchange Format TeDIF was developed in the framework of the TDCnet project – European Terminology Documentation Centre Network, co-funded by the EU Commission. The TeDIF format was developed with the purpose to establish a common format for bibliographical and factual data related to terminology.

TeDIF provides means to describe bibliographical data like literature (serials, monographs, articles, journals, theses, etc.) and term collections (printed dictionaries, glossaries, thesauri, classifications, terminology databases, etc.).

TeDIF is an SGML-based format (Standard Generalized markup Language, ISO 8879:1986) to describe and exchange data. Since TeDIF is also XML-compatible (Extended markup Language, subset of SGML), it is open to the newest developments in markup languages, the usage of Unicode, and an easier conversion to HTML and other formats.

In EuroTermBank project some minor but necessary additions for TeDIF were identified and implemented. Modifications include a possibility to multiply the fields describing the author and copyright holder according to the number of persons/organizations and the addition of fields for the indication of the languages of definitions and context information.

EuroTermBank contains description in TeDIF of 479 terminology resources identified in participating countries.:

- Estonian 119
- Hungarian 75
- Lithuanian 90
- Latvian 94
- Polish 101

After evaluation of these resources against quality criteria and completion of acquisition from respective IPR holders, 96 resources in 42 languages are currently included in EurotermBank.

EuroTermBank experience demonstrates applicability of TeDIF as a standard for terminology resource meta-data description and we recommend this format for other similar activities,

3.2 Application of TBX standard for data exchange

TBX (TermBase eXchange) is an open XML-based standard format for terminological data, created to facilitate interchange among termbases. This standard provides a number of benefits as TBX files can be imported into and exported from most software packages that include a terminological database[5]. For interoperability of terminological data, it is important to use open standards for data exchange. TBX as XML-based standard also provides platform-independent data exchange. It is intended to qualify as a TML (Terminology Markup Language), as defined in the TMF (Terminology Markup Framework) specified in ISO 16642:2003. In addition, TBX is intended to support the extraction and merging of information from other, non-TMF-compliant, formats, although these processes may involve some information loss. Besides TBX tags, the TBX format may include also meta-information tags, which allows including such information as HTML formatted data.

TBX standard is based on three ISO standards: ISO 12620, ISO 12200 and ISO 16642. ISO 12620 defines data categories to be used for terminological data storage either in digital or printed format. Terminological data categories described in this standard are divided into three large subgroups that contain more detailed data category sections:

- Term-related information
- Descriptive information
- Administrative information

As a standardized exchange format, TBX can be used as the interchange format between single system components. Moreover, it facilitates terminological information interchange among termbases with different data models, thus improving interoperability of terminological data globally.

The EuroTermBank system implements the TBX standard with required data categories to enable data exchange between different ETB modules, interoperability with external databases, data import/export and data store in the EuroTermBank internal database.

A list of required terminological data categories was created during the EuroTermBank project based on best practice research. Selected data categories were compared to data categories specified in ISO 12620 to verify their compatibility. As TBX standard defines XML-based format, it was possible to use only the required data categories and still be compatible with TBX standard.

Although TBX standard is mainly devised as an exchange format, in EuroTermBank it was also used for terminological data storage in the database, as terminology has specific characteristics that make it difficult to store such type of data:

- it has many optional data categories
- data categories frequently have no format restrictions
- size of some data categories is not predictable.

These problems were solved in EuroTermBank by storing data in the XML-based format defined in the TBX standard. This provided the following benefits:

- storage of all TBX data categories
- (virtually) no format and size limitations for data categories
- simple extensibility.

TBX standard is used also for data import and export to and from EuroTermBank database. All resources to be included in the portal's internal database are converted to TBX format. Source formats vary from printed paper resources to highly structured XML files. As TBX is also the storage format, there are no significant reasons for introducing another format. As TBX allows storage of all standardized data categories, it is possible to convert all resources to TBX format. Even if resources have resource specific data categories that are not included in the standard, it is possible to store these categories as supplemented XML tags without changing the physical data storage model.

TBX format is applied to data throughout the EuroTermBank system. Since TBX format is used trough all the resource life-cycle stages, it also ensures data consistency. Using an open standard is appropriate not only for EuroTermBank resource interoperability within the internal system, but also for communicating globally with external terminology databases. EuroTermBank system is designed to provide external systems with standardized data in TBX format and receive data from external systems in the very same way. There is no need to define a new framework either for

[5] *http://www.lisa.org/standards/tbx*

processing every single external data provider or for the data provided by the system.

In EuroTermBank system TBX standard enables data storage of all four terminological concept levels – entry level, language level, term level and word level (Vasiljevs, Schmitz, 2006). It also supports all data categories identified during the best practice research. All of 92 resources imported in EuroTermBank have been converted to TBX format without data loss, ensuring not only standard compliance, but also extensibility of the format.

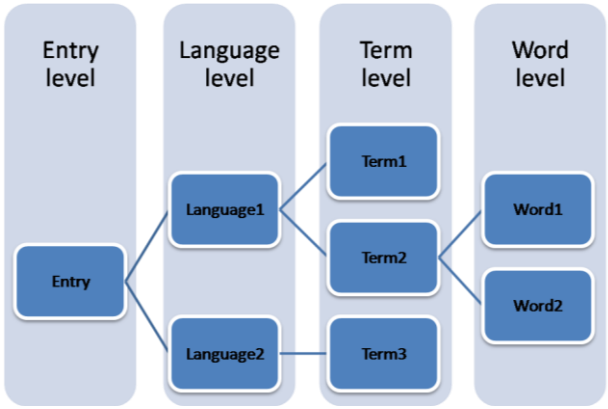


Fig. 1 EuroTermBank term data model

Using the TBX standard throughout the system provides data consistency as data are not converted either in the system internal modules or in the communications with external systems. From external systems what are already connected to the EuroTermBank system, one is directly providing data in TBX format. Other systems use proprietary exchange formats so conversion to TBX is applied before passing data to EuroTermBank. Furthermore, there are several systems that are on the way to use EuroTermBank system as the data source for terminology and communicate in the TBX standardized format.

Taken its strength in terminological data storage and exchange, TBX also has some weaknesses in data interoperability. TBX does not solve the problem of interoperability that originates from different application of data categories across term banks, for example, some data categories might be required in one termbase, while optional or not present in another one, or one and the same data category may appear on different levels of the entry structure. Also, there is no straightforward way for creating relations between terminological entries from different resources. Although technically it is possible, it is not part of the standard. The situation in creating relations between single resource entries is a bit better; a few types of relations – *broader*, *generic* and *related* – are defined there. However, these relations are limited and would be insufficient for creating more complex ontology structures.

4 Federation principle for unified access to independently maintained term banks

Federation is a new concept in linking portals and also data repositories, which goes far beyond the establishment of pointers or links, but reaches out to the level of semantic interoperability of data and data structures. Especially terminology and other kinds of structure content can be made to enable interoperability in the form of network(s) of federated databases (Galinski, 2007).

Despite application of international standards, terminology resources on the internet remain fragmented across diverse term banks and terminology projects. While it is clear that national or institutional terminology can be best identified in the terminology database of the respective institution, a number of user scenarios require consolidation on a multilingual and multinational scale. The goal of the EuroTermBank project is to not only centralize available terminology content in its database, but also act as a gateway that provides unified access to multiple remote terminology databases.
To ensure the viability of the federated system of terminology databases, inclusion of a termbank in this federated model requires it to be independently supported and maintained at both the institutional and technical levels.

Within EuroTermBank, the mechanism that enables federation of external databases is called interlinking. Interlinking an external database to EuroTermBank enables users to query the external database from EuroTermBank web interface. It is implemented by connecting to the external resource through a web service, ensuring platform-independent interoperable machine-to-machine interaction over a network. Communication is done using XML messages that follow the SOAP-standard, a protocol for exchanging XML-based messages over computer networks, normally using HTTP.

Several major external terminology databases are interlinked with EuroTermBank. An example of a national terminology database that is linked with EuroTermBank is the online databank of Latvian official terminology.

A number of challenges remain in implementation of the federated approach, such as ensuring the reliability of the sources or of the source data, ensuring a coordinated approach to change management, application or mapping of the potentially diverse subject field classification systems.

At the same time, the federated approach to terminology consolidation provides solution to at least one inherent challenge of all terminology banks – maintenance of terminology is done at the local or national level, and the changes at the local or national level become instantaneously available for integration with other federated resources.

5 Terminology entry compounding principle

This section describes EuroTermBank approach in unification of potentially matching terminology entries from different resources. Majority of terminology resources that are available in Eastern European countries are bilingual with a source language mostly being English. Much smaller number of resources is monolingual or has terms in three or more languages (Table1). This motivates us transform data representation from number of separate bilingual entries to unified multilingual record.

Entry languages	*Number of entries*	*Percentage from total*
monolingual	11230	2%
bilingual	398854	68%
3-lingual	45497	8%
4-lingual	69134	12%
5-lingual	48761	8%
>5-lingual	12216	2%

Table 1 Multilinguality of EuroTermBank source records

EuroTermBank data structure is modeled according to concept-oriented approach to terminology. Terminology entry denotes an abstract concept that has designations or terms as well as definitions in one or more languages. If terminology bank contains entries coming from different collections and designating the same concept we have an obvious interest to merge them into one unified multilingual entry. For example, if we have term pair EN *computer* – LV *dators* coming from Latvian IT terminology resource and another term pair EN *computer* – LT *kompiuteris* from Lithuanian IT terminology resource we may want to join these two into unified entry EN i – LV *dators* – LT *kompiuteris*. Such multilingual entry allows to get correspondence between language terms that are not directly available in any terminology resource (in our example new term pair LV *dators* – LT *kompiuteris*).

But merging entries just on the bases of matching term in one language that is common for these entries will lead to many erroneous term correspondences. Such problems are obvious due to the frequent ambiguity of terms among subject fields or much rarer cases of ambiguity in the context within one subject field. We can conclude that the only error-free method for merging entries is evaluating whether these entries denote the same concept. Unfortunately in practice it is often impossible or very expensive to make comparisons of cross-lingual terminology concepts. There is a lack of experts with sufficient knowledge of respective languages and subject fields. The task is considerably hindered by the fact that most terminology collections do not have definitions provided. In EuroTermBank, we propose a practical solution by introducing the terminology entry compounding approach. Entry compounding is an automated approach for matching terminology entries based on available data.

The most reliable indication for matching entries is having unique and unambiguous concept identifiers. The best example is terms from ISO terminology standards. These term entries have an identifier in the form [Standard_identifier].[term_number]. Accordingly, all national standards share the same identifier for corresponding entries and can be merged with a very high degree of reliability. Another case of unique internationally applied identification is the usage of Latin names in medicine and biology (with a number of exceptions with different Latin names designating the same concept). If there is no unique identification for concepts in collections, less precise matching criteria are used, namely, the English term and the subject field. English was chosen as the most popular language in term resources.

It is important to understand that entry compounding is a data representation method that does

not propose to create new terminology entries. It is a visualization aid that displays matching entries across collections in a consolidated way. Matches are determined by applying a number of criteria and as such cannot be error-free. Much like in machine translation environment, the user is prompted about potential incompatibilities and errors.

Entry compounding solves the problem of visual representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. At present, the EuroTermBank database contains over 585,711 term entries with more than 1,500,500 terms. When applying entry compounding, over 135,000 or 23% of entries get compounded. Hence entry compounding is a considerable aid for the user in finding the required term, for example, in the translation scenario between language pairs for which term equivalence is not established in existing collections. Unfortunately high recall rate lead also to relatively low precision although we currently do not have exact precision evaluation figures.

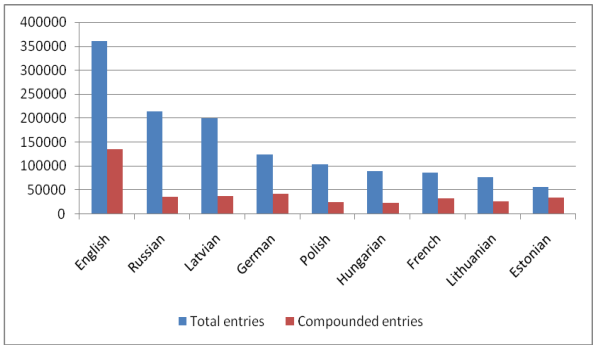


Table 2 Total and compounded entries per major languages of EuroTermbank.

At the same time, entry compounding may uncover incompatibilities and deficiencies across data and is therefore useful for further enhancement of the original data, but may be confusing for the immediate users.

The major source of problems for entry compounding lies in shortcomings of the subject field classification system and its application. In addition, entry compounding problems may occur due to different interpretations or errors while applying the classification system across all term collections, or the inherent differences across these collections. For example, errors may occur if a term within a subject field is used to denote several different concepts. This scenario contradicts to best practice methodology in terminology development, however, practice shows that existing term collections contain such deviant cases.

6 Terminology sharing

A number of emerging areas of activity are dependent on the availability and application of interoperable standards for terminology. One of them is terminology sharing – a new initiative that proposes sharing of non-confidential, non-competing and non-differentiating terminology across industrial companies and language service providers, with the goal to consolidate and promote accessibility to multilingual terminology per vertical industries (Rirdance, 2007). Terminology sharing involves returns from streamlined industry terminology that enhances customer satisfaction as well as, in future, common benefits from application of shared repositories for machine translation.

From the "ownership" perspective, there are two types of terminology resources:

- Public terminology resources
- Proprietary terminology resources

While public resources should be made accessible "by definition", there are a number of problems that hold back wide application of terminology sharing within the realm of public resources, such as IPR, institutional barriers and inertia, as well as technical incompatibilities.

Sharing of non-differentiating proprietary terminology involves taking appropriate measures to minimize the risk to the owner's intellectual capital and confidentiality. However, this is also a way of promoting and disseminating one's well-established terminology, possibly even to the level of *de facto* industry standard terminology.

7 Conclusions and Future Work

To summarize the most important points and lessons from the EuroTermBank project:

• observance and full application of standards in data consolidation is essential to interoperability and further applications of terminology data;

• entry compounding for representation of matching multilingual entries is applicable for creation of automatically formed multilingual terminology entries;

• federated approach in consolidation of resources enables distributed terminology to be ac-

cessible through a central gateway while it is maintained locally.

As a new type of terminology infrastructure providing access to diverse terminology resources, EuroTermbank can provide basis for further consolidation of terminology in Europe and beyond. Its rich and standards-based multilingual terminology resource collection, together with innovative instruments for analysis, can be used for research in terminology, lexicography, computational linguistics, as well as applied in computer-assisted translation systems.

Previous development phase was mostly concentrated on consolidation of large number of dispersed multilingual resources providing unified online accessibility. Further research and development will be concentrated on conceptualization of terminology data.

We are researching possibilities for elaboration of entry compounding using corpus analysis to evaluate compounding candidates. Further promising development directions are mechanisms for facilitating concept hierarchies and ontology integration.

Acknowledgements

Many thanks to colleagues in all EuroTermBank project partner organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland). EuroTermBank Consortium would also like to acknowledge and thank the European Union eContent program for supporting the EuroTermBank project as well as support from EU Social Fund.

References

Auksoriute A. (2006) *Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project.*, ISBN 9984-9133-4-1, Riga

Betz A.; Schmitz K.-D. (1999). The Terminology Documentation Interchange Format TeDIF. In: Terminology and Knowledge Engineering TKE'99, Innsbruck, Wien, pp. 782-792.

Galinski C. 2007. *New ideas on how to support terminology standardization projects*, eDITion, 1/2007.

Henriksen L., Povlsen C., Vasiljevs A. 2005. EuroTermBank – a Terminology Resource based on Best Practice. In *Proceedings of LREC 2006, the 5th International Conference on Language Resources and Evaluation*, Genoa, on CD-ROM , May 2006

Rirdance S. 2007. *IP vs. Customer Satisfaction: EuroTermBank and the Business Case for Terminology Sharing*, The Globalization Insider, LISA, 6/2007.

Vasiljevs A., Schmitz K.-D. 2006. *Collection, harmonization and dissemination of dispersed multilingual terminology resources in online terminology databank*, Proceedings of TSTT 2006, Third International Conference on Terminology, Standardization and Technology Transfer, Beijing, August 2006

Weissinger R. 2007. *ISO Concept Database* presentation, "Integrating Standards in Practice", 10th Open Forum on Metadata Registries, New York, July 2007.

Author Index