

Building Linguistic Resources in the Medical Domain

Pierre Zweigenbaum^{1,2}

¹CNRS, LIMSI, Orsay, France;

²INALCO, CRIM, Paris, France

Formerly at Assistance Publique – Hôpitaux de Paris and at INSERM U729, Paris

NACTEM Seminar, Feb. 16, 2007



Context

Antecedents

NLP and Medical Terminology

Assisted Acquisition of Lexical Resources

Monolingual Lexicon: Morphological Relations

Multilingual Lexicon: Finding Translations

Assisted Acquisition of Terminological Resources

Monolingual Terminology: Finding Variants

Multilingual Terminology: Finding Translations



Affiliations

Formerly at:

AP-HP: Chargé de mission,
Assistance publique –
Hôpitaux de Paris,
DSI/STIM (1984–2006)

Inserm: U194, U729
(1982–2006)

Currently at:

CNRS: Senior Researcher, LIMSI,
Orsay (2006–)

Inalco: Associate Professor,
CRIM (Multilingual
engineering research
center) (2003–2007)



Affiliations

Formerly at:

AP-HP: Chargé de mission,
Assistance publique –
Hôpitaux de Paris,
DSI/STIM (1984–2006)

Inserm: U194, U729
(1982–2006)

Currently at:

CNRS: Senior Researcher, LIMSI,
Orsay (2006–)

Inalco: Associate Professor,
CRIM (Multilingual
engineering research
center) (2003–2007)





Former Topics and Projects

| | | | |
|----------------------|------------------------------------|--------------------------------------|-----------------------------|
| <i>topic</i> | Access to contents of text corpora | Lexical and terminological resources | Text corpora for evaluation |
| <i>international</i> | NLPAD; <i>Menelas</i> ; DOME | Semantic Mining | |
| <i>national</i> | EQueR | UMLF; VUMeF | CLEF; EASy; EQueR; CESART |

All in the medical domain



NLP and Medical Terminology

Controlled indexing Detect that a document mentions a given topic / detect the occurrence of a term in a text

Assisted construction of termino-ontological resources

- Creation of new terminologies?
- Extension of existing terminologies: new variants, new concepts
- Translation of English language terminologies

Handling term variation

Identifying the occurrences of a term/concept under its different forms



NLP and Medical Terminology

Controlled indexing Detect that a document mentions a given topic / detect the occurrence of a term in a text

Assisted construction of termino-ontological resources

- Creation of new terminologies?
- Extension of existing terminologies: new variants, new concepts
- Translation of English language terminologies

Handling term variation

Identifying the occurrences of a term/concept under its different forms

Context

Antecedents

NLP and Medical Terminology

Assisted Acquisition of Lexical Resources

Monolingual Lexicon: Morphological Relations

Multilingual Lexicon: Finding Translations

Assisted Acquisition of Terminological Resources

Monolingual Terminology: Finding Variants

Multilingual Terminology: Finding Translations

Monolingual Lexicon: Morphological Relations

Target:

- monolingual lexicon
- morphological relations: complex words (affixation, compounding)

Sources:

- existing lexicons: issues of homogenisation and unification, project UMLF
- thesauri (terminologies): [▶ language retroengineering](#) (Grabar 1998–2002)
- corpus: [▶ thematic associations](#) are often grounded in morphological relations (Zweigenbaum 2003)

Also: automatic morphological analysis (Namer 2004) [Nancy]

Latent Morphological Relations in a Thesaurus

SNOMED: French Microglossary for Pathology (12,555 terms, 2,344 series of synonyms) ▶ RIAO 2000

| <i>Code</i> | <i>Class</i> | <i>Term</i> |
|-------------|--------------|--------------------------------|
| D0-10430 | 01 | pemphigoïde, SAI |
| D0-10430 | 02 | pemphigus bénin, SAI |
| D6-50530 | 01 | déficit en galactose épimérase |
| D6-50530 | 02 | galactosémie type III |
| F-C6000 | 01 | cellule immunitaire |
| F-C6000 | 02 | immunocyte |
| F-D0650 | 01 | fonction hémostatique |
| F-D0650 | 02 | hémostase |
| M-02712 | 01 | consistance anormalement dure |
| M-02712 | 02 | durcissement |
| M-02712 | 02 | durci |
| M-35300 | 01 | embole |
| M-35300 | 02 | embolie |
| M-35300 | 05 | embolique |

Latent Morphological Relations in a Corpus

Find the relational adjective for a given noun ▶ AMIA 2003

apophyse

appendice

cardia

cotyloïde

cristallin

diaphyse

éosinophile

hippocampe

intima

jambe

lysosome

macrophage

mastocyte

myomètre

métatarse

néphron

olécrâne

paramètre

plasma

apophysaire

appendiculaire

cardial

cotyloïdien

cristallinien

diaphysaire

éosinophilique

hippocampique

intimal

jambier

lysosomal

macrophagique

mastocyttaire

myométrial

métatarsien

néphronique

olécrânien

paramétrial

plasmatique

- Input: 376 nouns in the domain of anatomy (SNOMED)
- Tagged corpus (5 Mwords)
- **Lexical consequence of thematic coherence**: words of a given morphological family are often used in the same paragraph
- Identified 150 relational adjectives (precision = 91%) among which **22 were completely absent from this nomenclature**

Latent Morphological Relations in a Corpus

Find the relational adjective for a given noun ▶ AMIA 2003

| | |
|--------------------|-----------------------|
| <i>apophyse</i> | <i>apophysaire</i> |
| <i>appendice</i> | <i>appendiculaire</i> |
| <i>cardia</i> | <i>cardial</i> |
| <i>cotyloïde</i> | <i>cotyloïdien</i> |
| <i>cristallin</i> | <i>cristallinien</i> |
| <i>diaphyse</i> | <i>diaphysaire</i> |
| <i>éosinophile</i> | <i>éosinophilique</i> |
| <i>hippocampe</i> | <i>hippocampique</i> |
| <i>intima</i> | <i>intimal</i> |
| <i>jambe</i> | <i>jambier</i> |
| <i>lysosome</i> | <i>lysosomal</i> |
| <i>macrophage</i> | <i>macrophagique</i> |
| <i>mastocyte</i> | <i>mastocytaire</i> |
| <i>myomètre</i> | <i>myométrial</i> |
| <i>métatarse</i> | <i>métatarsien</i> |
| <i>néphron</i> | <i>néphronique</i> |
| <i>olécrâne</i> | <i>olécrânien</i> |
| <i>paramètre</i> | <i>paramétrial</i> |
| <i>plasma</i> | <i>plasmatique</i> |

- Input: 376 nouns in the domain of anatomy (SNOMED)
- Tagged corpus (5 Mwords)
- **Lexical consequence of thematic coherence**: words of a given morphological family are often used in the same paragraph
- Identified 150 relational adjectives (precision = 91%) among which **22 were completely absent from this nomenclature**

Context

Antecedents

NLP and Medical Terminology

Assisted Acquisition of Lexical Resources

Monolingual Lexicon: Morphological Relations

Multilingual Lexicon: Finding Translations

Assisted Acquisition of Terminological Resources

Monolingual Terminology: Finding Variants

Multilingual Terminology: Finding Translations

Multilingual Lexicon: Finding Translations

Target:

- multilingual lexicon
- translation relations between words
- project: multilingual medical lexicon, «Semantic Mining»
European Network of Excellence

Sources:

- existing lexicons: reuse, [▶ training](#) (Claveau 2005), alignment by [▶ decomposition](#) (Markó 2006)
- multilingual thesauri (terminologies): [▶ word alignment](#) (Baud 1998, Nyström 2006)
- parallel bilingual corpora: [▶ word alignment](#) (Deléger 2006)
- comparable bilingual corpora: [▶ multilingual distributional analysis](#) (Chiao 2002–2004)



Translation by Transduction

(Vincent Claveau, 2005) [▶ AIME2005](#)

Source: partial bilingual lexicon which contains “graphically similar” word pairs

- The partial bilingual lexicon serves as training set
- Learns a transducer which converts a word from one language to the other
- Applies transducer to new words

Translation by Decomposition into “Subwords”

(Kornél Markó *et al.*, 2006) [Freiburg] ▸ LREC2006

Source: independent monolingual lexicons

1. In each monolingual lexicon: decomposition of input words into “subwords”
2. Alignment of decompositions

▸ ... extension of DériF to French neoclassical compounds
(Namer 2004–2006) [Nancy]

▸ ... extension of DériF to English neoclassical compounds
(Deléger 2007)

Word Alignment in Parallel Terms

- Source: multilingual terminology
 - International Classification of Diseases (ICD-10)
- Rationale: analogy with word alignment in parallel sentences

(Baud *et al.*, 1998) [Geneva], (Nyström *et al.*, 2006) [Linköping]

- Each term plays the role of a sentence
- Application of a word alignment algorithm

Parallel Terms: ICD-10

| <i>code</i> | <i>English</i> | <i>French</i> |
|-------------|-------------------------|------------------------|
| A20 | Plague | Peste |
| A20.0 | Bubonic plague | Peste bubonique |
| A20.1 | Cellulocutaneous plague | Peste cutanée |
| A20.2 | Pneumonic plague | Peste pulmonaire |
| A20.3 | Plague meningitis | Peste méningée |
| A20.7 | Septicaemic plague | Peste septicémique |
| A20.8 | Other forms of plague | Autres formes de peste |
| A20.9 | Plague, unspecified | Peste, sans précision |

Word Alignment in Parallel Corpora

(Deléger *et al.*, 2006) ▶ LREC2006

Source: parallel corpora (Health Canada – Santé Canada)

- Alignment of documents
- Alignment of sentences (GMA, Melamed 1999)
- Alignment of words (I*Tools, Ahrenberg *et al.*, 2003)

Word Alignment In Comparable Corpora

(Chiao *et al.*, 2002–2004) ▶ comparable

Sources: comparable corpora CISMef – CliniWeb [C4];
partial bilingual lexicon

- Constitution of both corpora: medical Web catalogs
- Distributional analysis in each corpus
 - The distributional profile of each word is represented by a context vector
- Cross-lingual distributional similarity:
 - Convert “source” distributional profiles into “target” profiles through partial bilingual lexicon
 - Compare source word profile to target words profiles
 - Rank target words by decreasing similarity



Translation by Comparable Corpora: Examples

| French | <i>First proposed translations</i> |
|-------------|--|
| carence | <i>deficiency, vitamin, folate, iron, low, zinc</i> |
| angoisse | <i>anxiety, depression, panic, attack, agitation</i> |
| nécrose | <i>necrosis, chronic, renal, inflammation, infraction</i> |
| gène | <i>gene, mutation, protein, chromosome, recessive</i> |
| sclérose | <i>sclerosis, sep, lateral, passe, poliovirus</i> |
| abcès | <i>abscess, perforation, rupture, visible, invasive, impose</i> |
| hépatite | <i>hepatitis, infection, virus, invasive, immunodeficiency</i> |
| aorte | <i>aortic, carotid, aneurysm, aorta, artery, coronary, left</i> |
| acide | <i>amino, protein, acid, fatty, deficiency, folic, iron, zinc, substance</i> |
| plasmatique | <i>plasma, serum, high, low, increased, lithium</i> |

Context

Antecedents

NLP and Medical Terminology

Assisted Acquisition of Lexical Resources

Monolingual Lexicon: Morphological Relations

Multilingual Lexicon: Finding Translations

Assisted Acquisition of Terminological Resources

Monolingual Terminology: Finding Variants

Multilingual Terminology: Finding Translations

Monolingual Terminology: Finding Variants

Target:

- more term variants in a monolingual terminology
- project: VUMeF, increase the French part of the UMLS Metathesaurus

Sources:

- existing terms (MeSH thesaurus)
- morphological relations (relational adjectives)
- domain corpus: attestations

EQueR Corpus: 19 M words ; UMLF relational adjectives; Syntex parser (Bourigault et al. 2003)

- ▷ Project VUMeF, 2004–2006; coll. CISMef, ERSS, ATILF, VIDAL, ...

Terminology Extension: Term Variants

- Variants of N ($Prep$) $N_1 \leftrightarrow N A_1$ MeSH terms
- 309 new variants proposed, precision ~ 54%

| # occurrences | | | | $N(Prep) N_1$ | | | NA_1 |
|---------------|-------|-------|-------|---------------|---------------|--------------|--------------------------|
| n_t | a_t | n_m | a_m | | | | |
| 1 | 121 | 1 | 0 | cancer | | bronche | cancer bronchique |
| 1 | 44 | 1 | 0 | évaluation | | technologie | évaluation technologique |
| 1 | 19 | 1 | 0 | tumeur | | thymus | tumeur thymique |
| 1 | 2 | 1 | 0 | médecine | | reproduction | médecine reproductif |
| 1 | 1 | 1 | 0 | cancer | | thorax | cancer thoracique |
| 1 | 1 | 1 | 0 | professionnel | en | éducation | professionnel éducatif |
| 1 | 1 | 0 | 1 | besoin | en matière de | nutrition | besoin nutritif |
| 1 | 1 | 0 | 1 | chambre | de | malade | chambre malade |
| 1 | 1 | 0 | 1 | diarrhée | de | bovin | diarrhée bovin |
| 1 | 1 | 0 | 1 | fragment | de | peptide | fragment peptidique |
| 1 | 1 | 0 | 1 | infection | dans | oeil | infection ophtalmique |
| 1 | 1 | 0 | 1 | médicament | pour | animal | médicament animal |
| 1 | 1 | 0 | 1 | toxine | de | choléra | toxine cholérique |
| 1 | 2 | 0 | 2 | cellule | de | estomac | cellule gastrique |
| 1 | 3 | 0 | 3 | résistance | de | champignon | résistance fongique |

Context

Antecedents

NLP and Medical Terminology

Assisted Acquisition of Lexical Resources

Monolingual Lexicon: Morphological Relations

Multilingual Lexicon: Finding Translations

Assisted Acquisition of Terminological Resources

Monolingual Terminology: Finding Variants

Multilingual Terminology: Finding Translations

Multilingual Terminology: Finding Translations

Target:

- French translations of English terms
- (→ more term variants in a monolingual terminology)
- project: VUMeF, increase the French part in the UMLS Metathesaurus

Sources:

- English terms
 - MeSH thesaurus; SNOMED CT nomenclature
- Health Canada – Santé Canada parallel corpus



Translation by Parallel Corpora: Examples

| American MeSH | French MeSH | Health Canada parallel corpus |
|------------------------|-------------------------------|-------------------------------|
| cheek bones | Apophyse zygomatique | pommettes des joues |
| Death | Mort | Décès |
| Asian | Groupe d'ascendance asiatique | Asiatique |
| Atomic Energy | Énergie nucléaire | énergie atomique |
| vegetables | Plantes potagères | légumes |
| fertility rate | Taux natalité | taux de fécondité |
| Infant Mortality | Mortalité du nourrisson | Mortalité infantile |
| dizziness | Sensation vertigineuse | sensation de vertige |
| vibration | Vibration | vibratoire |
| vulnerable populations | Populations vulnérables | populations à risque |
| West Nile virus | Virus du Nil occidental | VNO |
| wounds | Plaies et traumatismes | blessures |
| construction materials | Matériaux construction | matériaux de construction |
| birth weight | Poids naissance | Poids à la naissance |