HIMERA CORPUS

1 BASIC INFORMATION

1.1 Corpus composition

The HIMERA corpus (HIstory of Medicine CoRpus Annotation) consists of:

- 35 articles from the archives of the British Medical Journal (BMJ)
- 4 extracts of reports from the archives of the London Area Medical Officer of Health (MOH) reports

The articles and extracts are drawn from 4 key decades in the long histories of these are archives, i.e. the 1850s, 1890s, 1920s and 1960s. The articles and extracts were mostly selected according to their mentions of lung diseases.

1.2 Representation of the corpus (flat files, database, markup)

The corpus consists of a set of plain text files (with *.txt* extensions), accompanied by files containing stand-off annotations (with *.a1* and *.a2* extensions)

1.3 Character encoding

The characters are UTF8 encoded.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person

Name: Sophia Ananiadou
Address: Manchester Institute of Biotechnology,131 Princess Street, Manchester M1 7DN, IK
Affiliation: National Centre for Text Mining, School of Computer Science, University of Manchester
Position: Professor
Telephone: +44 161 306 3092
Fax: +44 161 306 5201
e-mail: Sophia.Ananiadou@manchester.ac.uk

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the META-SHARE platform as an archive.
The corpus can also be visualised online at:
http://www.nactem.ac.uk/brat/#/HIMERA/

2.3 Copyright statement and information on IPR

The resource is licensed under a Creative Commons Attribution licence (CC-BY). If you use the resource, please attribute:

a) The National Centre for Text Mining (NaCTeM), who created the annotations. Please also cite the following article:

Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys and Sophia Ananiadou (2015). Text Mining the History of Medicine. *PLOS ONE*.

b) The British Medical Journal (BMJ), who kindly consented to the use of the 35 articles from the BMJ archive.

c) The Wellcome Library, who made available the MOH reports. These reports are also licenced under a licensed under a Creative Commons Attribution licence (CC-BY).

3 TECHNICAL INFORMATION

3.1 Directories and files

The archive contains the directory *HIMERA_Corpus*. It contains the following sub-directories and files:

- *BMJ* – directory containing the articles and associated annotations from the BMJ archive. It contains four sub-directories, corresponding to the different decades from which the articles are drawn, i.e.:
  - *1850s*
  - *1890s*
  - *1920s*
  - *1960s*

  Each sub-directory contains plain text files (.txt) and associated files containing stand-off annotations (with .a1 and .a2 extensions). These standoff files follow the format of the corpora created in the context of the BioNLP 2013 Shared Task, as described below.
- *MOH* – directory containing plain text files (.txt) and associated annotation files (.a1 and .a2) for 4 extracts from MOH reports, one from each of the following decades: 1850s, 1890s, 1920s and 1960s.
  The standoff files follow the format of the corpora created in the context of the BioNLP 2013 Shared Task, as described below.
- *HIMERA_README.txt* – provides a range of information about the annotated corpus including the annotation scheme, composition of the corpus and the format of the annotation files.

- *HIMERA_licence.txt* – Provides information about the licence applied to the corpus.

3.2 Data structure of an entry

This is not relevant as the corpus is a set of text files.

3.3 Resource size (nmb. of tokens, MB occupied on disk)

The corpus contains approximately 70,000 tokens. It requires approximately 1.2 MB of disk space.

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

This corpus is a monolingual, annotated corpus.

4.2 The natural language(s) of the corpus

The language of the corpus is English.

4. 3 Domain(s)/register(s) of the corpus

The corpus contains articles and excerpts of published historical medical text

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

Each text file contains one sentence per line. Each sentence is manually annotated with semantic information that is relevant to the study of medical history and public health. Specifically, two levels of semantic information have been annotated, based on extensive discussions with medical historians:

- Named entities belonging to seven different categories of medical and/or historical significance.
- Two different event types, which encode relationships amongst entities and other events.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

ENTITY ANNOTATIONS

Brief definitions of the seven entity types annotated in HIMERA are provided in Table 1

Table 1. Entity types annotated in HIMERA

Entity Type	Description	Examples
Condition	Medical condition/ailment	*phthisis, bronchitis, typhus*
Sign_or_ Symptom	Altered physical appearance/behaviour as probable result of injury/condition	*cough, pain, rise in temperature, swollen*
Anatomical	Entity forming part of human body, including substances and abnormal alterations to bodily structures	*lung, lobe, sputum, fibroid*
Subject	Individual or group under discussion	*children, asthma patients, those with negative reactions to tuberculin*
Therapeutic_or_ Investigational	Treatment/intervention administered to combat condition (including diet/foodstuffs), or substance, medium or procedure used in investigational medical or public health context	*atrophine sulphate, generous diet, change of air, lobectomy*
Biological Entity	Living entity not part of human body, including microorganisms, animals and insects	*tubercle bacilli, mould, guinea-pig, flea*
Environmental	Environmental factor relevant to incidence/prevention/control/treatment of condition. Includes climatic conditions, foodstuffs, infrastructure, household items or occupations whose environmental factors are mentioned	*humidity, high mountain climates, infected milk, linen, drains, sewers, dusty occupations*

EVENT ANNOTATIONS

Events annotations consist of the following:

- **A trigger**, i.e. a word or phrase that characterises the event. Frequently a verb (e.g., "caused") or a nominalisation (e.g., "affect").
- **A semantic type** for the event.
- **A set of participants** (entities or other events), which contribute towards the description of the event. Participants may be entities, or previously annotated events. Each participant is assigned a semantic role, according to the type of information that it contributes towards the overall description of the event.
- If an event is negated, it is assigned a **negation attribute, and an associated negation cue** (i.e. a word or phrase indicating the negation) may be identified.

Definitions of the two event types annotated in HIMERA, along with their participant types/semantic roles, are provided in Table 2

Table 2. Event types annotated in HIMERA

Event Type	Description	Possible Participants
Affect	A (previously existing) entity or event is affected, infected, undergoes change or is transformed, possibly by another entity or event	**Cause:** Cause of the affection **Target:** Entity or event affected **Subj:** Individual or group affected
Causality	An entity or event results in the manifestation of a (previously non-existing) entity or event	**Cause:** Cause of the manifestation **Result:** New entity or event that manifests itself **Subj:** Individual or group associated with the event

4.4.3Alignment information (if the corpus contains aligned documents: level of alignment, how it was achieved)

Not applicable – this is a monolingual corpus.

4.4.4 Attributes and their values (if annotated)

Annotations are encoded in the BioNLP Shared Task 2013 format (http://2013.bionlp-st.org/file-formats). Based on this format, there are two annotation files associated with each text file:

- **a1 files** - encode information about entity annotations (and negation cues)
- **a2 files** - encode information about event annotations

A1 FILES

In a1 files, each line provides information about a single entity. Examples are shown below:

```
T1      Condition 50 84     acute suffocative pulmonary oedema
T2      Sign_or_Symptom 6612 6618;6630 6642     mitral incompetence
T9      Negation_Cue 5609 5612    not
```

Each line consists of:

- **A unique id for the entity**. By convention, this starts with *T*, followed by a numerical value.
- **A TAB character.**
- **The entity type assigned to the annotation** (or the label *Negation_Cue*, in the case that the annotated span corresponds to a word or phrase indicating the negation of an event).
- **The character-based offsets of the entity annotation in the corresponding text file.** There are two formats for the offsets, depending on whether the annotated span consists of a single, continuous span or a discontinuous span, consisting of multiple, connected spans. As an example of cases where a discontinuous annotation is needed, consider the text span *gouty or rheumatic bronchitis*. In this span, there are two conditions mentioned, i.e., *gouty bronchitis* and *rheumatic bronchitis*, although the word *bronchitis* appears only once. In order to annotate *gouty bronchitis*, it is necessary to annotate the words *gouty* and *bronchitis*, and to link them together.
    - For continuous spans (as in the first example line above), there

are two offsets, corresponding to the start and end offsets of the span. The first offset is separated by a space from the entity type label, and there is a space between the start and end offsets
    - For discontinuous spans (as in the second example line above), there are two or more pairs of start and end offsets, each separated by a semi-colon. Each pair of offsets corresponds to a part of the complete annotated span.
- **Another TAB character**
- **The text covered by the annotated span in the corresponding text file.**

A2 FILES

There are three different formats of lines in a2 files, as shown in the example below:

```
T10        Affect 3347 3358       gave relief
E11        Affect:T10 Cause:T63 Subj:T61 Cue:T43
E8         Causality:T11 Result:T240 Result2:T254 Cause:T241
M14        Negation E11
```

The formats of the lines are as follows:

- **Lines starting with *T*** - These correspond to event trigger spans and have exactly the same format as the lines in the a1 files, except that the semantic labels correspond to the relevant event type (i.e., either *Causality* or *Affect*). As with entity annotations, the spans may be discontinuous.
- **Lines starting with *E*** - These correspond to event annotations. They consist of the following parts:
    - **A unique id for the event**. By convention, this starts with an E, followed by a numerical value.
    - **A TAB character**.
    - **The semantic type assigned to the event, followed by a colon, and the ID assigned to the event trigger span**
    - **A sequence of pairs of the format [Label]:[ID], separated by spaces**. Each pair corresponds either to an event participant, in which case the [Label] part of the pair is the semantic role assigned to the participant, or to a negation cue, in which case the [Label] part has the value Cue. The [ID] part may start with a *T*, in which case it corresponds to an entity annotation in the associated a1 file, or, it may start with an *E*, in which case the participant corresponds to another event listed within the same a2 file. If more than one participant is assigned the same semantic role, then for the second and subsequent participants, a number is appended to the semantic

role label, e.g., *Result2* for the second participant assigned the Result role, as in the third example line above.

- **Lines starting with *M*** - These correspond to attributes or modifications assigned to events. In the case of HIMERA, the only such modification possible is for an event to be negated. These lines consist of the following parts:
- **A unique id for the negation**. By convention, this starts with an *M*, followed by a numerical value.
- **A TAB character**.
- **The label *Negation*, followed by a space and then the ID of the event that is negated**

4.5 Intended application of the corpus

HIMERA is intended to provide the means to train and evaluate text mining (TM) tools that are able to recognise relevant entities and relationships (or events) that hold between them, in a range of types of published medical documents, dating from the mid 19th century onwards.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

Approximately a quarter of the corpus was double annotated to allow inter-annotator agreement (IAA) rates to be calculated, and to ensure the consistency of the annotations. We calculated IAA in terms of F-Score, and found that high levels of agreement were achieved. For exact span matches (i.e., where the start and end of the annotated text spans chosen by both annotators must match exactly), the IAA was 0.80 F-Score. For relaxed matches, where it is sufficient for the annotations to include some common parts, the IAA was 0.86 F-Score.

5 RELEVANT REFERENCES AND OTHER INFORMATION

The HIMERA corpus is introduced in the following article:

Paul Thompson, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys and Sophia Ananiadou (2015). Text Mining the History of Medicine. *PLOS ONE*.