Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 113−135

www.elsevier.com/locate/csl

A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation[☆]

Scott Piao[a,*], Fraser Dallachy[b], Alistair Baron[a], Jane Demmen[a], Steve Wattam[a], Philip Durkin[c], James McCracken[c], Paul Rayson[a], Marc Alexander[b]

[a] *Lancaster University, Lancaster LA1 4WA, United Kingdom*
[b] *University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[c] *Oxford University Press, Oxford OX2 6DP, United Kingdom*

Received 22 July 2016; received in revised form 5 February 2017; accepted 30 April 2017
Available online 17 May 2017

Abstract

Automatic extraction and analysis of meaning-related information from natural language data has been an important issue in a number of research areas, such as natural language processing (NLP), text mining, corpus linguistics, and data science. An important aspect of such information extraction and analysis is the semantic annotation of language data using a semantic tagger. In practice, various semantic annotation tools have been designed to carry out different levels of semantic annotation, such as topics of documents, semantic role labeling, named entities or events. Currently, the majority of existing semantic annotation tools identify and tag partial core semantic information in language data, but they tend to be applicable only for modern language corpora. While such semantic analyzers have proven useful for various purposes, a semantic annotation tool that is capable of annotating deep semantic senses of all lexical units, or all-words tagging, is still desirable for a deep, comprehensive semantic analysis of language data. With large-scale digitization efforts underway, delivering historical corpora with texts dating from the last 400 years, a particularly challenging aspect is the need to adapt the annotation in the face of significant word meaning change over time. In this paper, we report on the development of a new semantic tagger (the Historical Thesaurus Semantic Tagger), and discuss challenging issues we faced in this work. This new semantic tagger is built on existing NLP tools and incorporates a large-scale historical English thesaurus linked to the Oxford English Dictionary. Employing contextual disambiguation algorithms, this tool is capable of annotating lexical units with a historically-valid highly fine-grained semantic categorization scheme that contains about 225,000 semantic concepts and 4,033 thematic semantic categories. In terms of novelty, it is adapted for processing historical English data, with rich information about historical usage of words and a spelling variant normalizer for historical forms of English. Furthermore, it is able to make use of knowledge about the publication date of a text to adapt its output. In our evaluation, the system achieved encouraging accuracies ranging from 77.12% to 91.08% on individual test texts. Applying time-sensitive methods improved results by as much as 3.54% and by 1.72% on average.

Keywords: Semantic annotation; Natural language processing; Historical thesaurus; Semantic lexicon; Corpus annotation; Language technology

[☆] This paper has been recommended for acceptance by Pascale fung.

[*] Corresponding author.

E-mail address: s.piao@lancaster.ac.uk (S. Piao), fraser.dallachy@glasgow.ac.uk (F. Dallachy), a.baron@lancaster.ac.uk (A. Baron), j.e.demmen1@lancaster.ac.uk (J. Demmen), s.wattam@lancaster.ac.uk (S. Wattam), philip.durkin@oup.com (P. Durkin), james.mccracken@oup.com (J. McCracken), p.rayson@lancaster.ac.uk (P. Rayson), marc.alexander@glasgow.ac.uk (M. Alexander).

http://dx.doi.org/10.1016/j.csl.2017.04.010

1. Introduction[1]

Semantic analysis of natural language data is a relevant task for a wide range of research areas and practical applications, such as natural language processing, text mining, corpus linguistics and data science. Numerous semantic annotation tools have been developed to carry out various levels of semantic analysis, such as document topics, named entities, temporal information, and so on. For example, some tools are designed to identify the topic or themes of given texts (Allan, 2012), and some are designed to extract specific partial information, such as types of named entities, categories of relations between the specific named entities, and/or types of events (Miwa et al., 2012; Rizzo and Troncy, 2012; Weston et al., 2013). Another group of semantic annotation tools are designed to identify semantic categories of all lexical units based on a given classification scheme, which can support a deep comprehensive semantic information analysis and extraction from language data. The latter task entails richer semantic lexical resources and a deeper level of sense disambiguation, and hence presents tough challenges. Our work presented in this paper addresses the issue of a semantically rich text analytical system.

Over recent years, various semantic lexical resources and semantic annotation tools have been developed, such as EuroWordNet (Vossen, 1998) and the UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System (USAS) (Rayson et al., 2004), and they have played an important role in developing intelligent natural language processing (NLP) and Human language technology (HLT) systems. For example, the USAS semantic tagger has been applied in a variety of studies, including empirical language studies at the semantic level (Klebanov et al., 2008; Ooi et al., 2007; Potts and Baker, 2013; Rayson et al., 2004), studies in information technology (Doherty et al., 2006; Nakano et al., 2005; Volk et al., 2002), software engineering (Chitchyan et al., 2006; Taiani et al., 2008) and others (Balossi, 2014; Gacitua et al., 2008; Hancock et al., 2013; Markowitz and Hancock, 2014; Semino et al., 2015).

In this paper, we present our work in designing, developing and evaluating the accuracy of a new semantic tagger: the "Historical-Thesaurus-based Semantic Tagger" (henceforth HTST). The purpose of this tool is to annotate all lexical units of texts with a fine-grained semantic categorization scheme provided by a very large-scale and high-quality English historical thesaurus (Kay et al., 2016 [2009]) (detailed further in the next section).

2. Related work

In recent years, researchers have devoted a great deal of effort to the development of various semantic annotation tools of natural language data. In particular, various lexical knowledge bases have been used to assign semantic concepts and categories to words and other types of lexical units in text. For example, WordNet is widely used for such a purpose, as demonstrated by the collection of WordNet Sense annotated corpora at the website http://globalwordnet.org/wordnet-annotated-corpora (last accessed 19 April 2016). A similar approach has been used for developing a more semantic field oriented semantic tagger, USAS, at UCREL (Lancaster University, UK; http://ucrel.lancs.ac.uk/usas), which is based on semantic lexicons containing lexical units classified with a set of pre-defined coarse-grained semantic fields rather than grouped by fine-grained word senses as in WordNet.

A significant amount of effort has been dedicated in previous research to word sense disambiguation, in particular in the SensEval series of events (Evaluation Exercises for the Semantic Analysis of Text; http://www.senseval.org), and more recently this has widened out (in SemEval and *SEM) to encompass other elements of computational analysis of meaning. Although, in some cases these do use existing sense inventories (e.g. BabelNet), generally the sense inventory is induced or clustered from a training set. Corpus-based distributional semantic models and word embeddings are now proving a very popular approach but generally still conflate different meanings of words under a single vector representation. In some works (Iacobacci et al., 2015), this limitation is starting to be addressed, but so far no research has been able to leverage meaning change over time, and this is obviously key for semantically annotating

[1] Abbreviations: CLAWS=Constituent Likelihood Automatic Word-tagging System; EEBO=Early English Books Online; EModE=Early Modern English; GATE=General Architecture for Text Engineering; HTST=Historical Thesaurus Semantic Tagger; MWE=MultiWord Expression; NLP=Natural Language Processing; OE=Old English; OED=Oxford English Dictionary; POS=Part-of-Speech; SAMUELS=Semantic Annotation and Mark-up for Enhancing Lexical Searches; UCREL=University Centre for Computer Corpus Research on Language; USAS=UCREL Semantic Analysis System; VARD=Variant Detector Software.

historical corpora where modern taxonomies and sense clusters would fail to capture historically valid word meanings.

Besides the all-words annotation tools mentioned above, which attempt to assign semantic categories to every word and lexical unit, numerous semantic annotation tools have been developed aiming to identify and assign certain types of semantic information requested by specific tasks to part of the lexical units or text segments, such as types of named entities and events (Named Entity Recognition), relations between entities (Semantic Role Labeling), attributes of product names (Sentiment Analysis), content analysis, and temporal information of events. A typical software framework developed for such a purpose is GATE (General Architecture for Text Engineering; Cunningham et al., 2011), which provides semantic annotation functionalities for various levels of semantic annotations. However, our focus in this paper is on fine-grained (deep) word sense disambiguation relative to existing historically sensitive semantic taxonomies.

Directly related to our paper is the development of the *Historical Thesaurus of English* (hereafter HT; Kay et al., 2016 [2009]) carried out at the University of Glasgow, UK. The HT is the result of over four decades' manual compilation by experts, which classifies the recorded vocabulary of English from the Old English period to the present day in a comprehensive semantic structure. The semantic classification is based primarily on a systematic analysis of the content of the Oxford English Dictionary, with other content from additional dictionaries of English. To this end, words are arranged into categories by the concepts they express, with successive subdivision of these categories delineating ever more precise sub-concepts within a concept. (For further details, see the HT website: http://www.glasgow.ac.uk/thesaurus.)

The HT database consists of two datasets: lexicon dataset and category dataset, which are linked via HT category numerical IDs. For example, in the sample HT entries below, the word "mother" in the HT lexicon dataset has a HT category ID "6959"(second number in the entry), along with various information. This ID is used to link to a definition entry in the HT category dataset (the first number in the lexicon sample entry). The category entry contains various information such as the HT category tag code "01.01.10.12.02.03" and heading "mater". For most words, there are multiple lexicon entries which link the words to multiple HT categories.

[Sample HT entries]
Lexicon entry:
"22,028";"6959";"mother";;"mother";"c1391";"1391";"1391";;;"c";"1391";"0";;;;;"0";"0";;;;"0";"0";;;;"321";"0"
Category entry:
"6959";"01";"01";"10";"12";"02";"03";;"02.01";"n";"mater";"sub.6.2";"01.01.10.12.02.03";"A26";"123"

The creation of the HT was instituted by Michael Samuels at Glasgow in 1965 and completed under the supervision of Christian Kay in 2009, at which point it was both released online (at www.glasgow.ac.uk/thesaurus; last accessed 19 April 2016) and printed by Oxford University Press under the title *Historical Thesaurus of the Oxford English Dictionary*. Old English (OE) vocabulary is not present in the *Oxford English Dictionary* (OED) except where a direct reflex of an OE form is found in the present day language, and so to fill this gap, *A Thesaurus of Old English* (Roberts et al., 1995) was produced as a pilot project to the main thesaurus, acting as a proof of concept and a further input source for its parent project.

The HTST tagger that we report on in this paper falls under the category of all-words tagging. It is based on an English thesaurus knowledge base, and hence employs a highly finely-grained semantic classification scheme that has not previously been incorporated into any language semantic annotation tools. Our work therefore extends the capability of the existing semantic taggers in terms of both the depth and scale of annotation of language data. Since the HT entries are linked directly to OED senses, we are also able to train a tagger, for the first time, on the wealth of information about each sense encoded in the definitions and example sentences in the OED database.

3. Structure of Historical Thesaurus entries

The HT database provides rich information about the usage of words. During the HTST development, the information concerning the disambiguation of the HT semantic categories of the words was selected and incorporated in the system. Such information includes HT ID (numerical codes used to index words in HT), HT semantic category ID (numerical codes representing semantic categories), part of speech (POS) information, and the time period during which the lexeme with that word sense is active in English, as well as headwords that are used to define the HT semantic categories. Fig. 1 outlines the above mentioned information network of the HT lexemes.

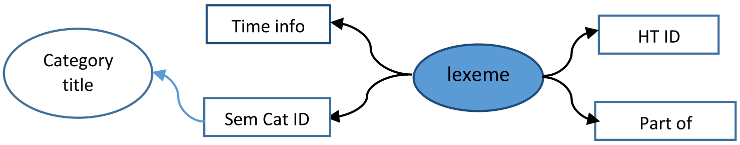


Fig. 1. Main information of HT entries used in HTST system.

The HT semantic categories are organized in a hierarchical structure, with each category being assigned a label and numeric code. The top level of classification is divided into ‘The External World’, ‘The Mental World’, and ‘The Social World’, prefixed with numeric codes 01, 02 and 03, respectively. At the next level of the hierarchy are 377 major categories, such as ‘The earth’, ‘Language’, and ‘Armed hostility’. The numeric codes extend according to the number of levels in the hierarchy in which the category is found. For example, ‘The Mental World’ is coded 02; within that, ‘Language’ is coded 02.07; at the next level down, ‘A language’ is coded 02.07.01, and so forth, down to a maximum of seven levels. Each category level may also contain subcategories used to classify lexis which provides finer details about the concept in question, but which does not constitute a new category in its own right. Altogether, the 2016 version of the HT contains 225,131 semantic divisions, which are used to classify 793,742 word forms. In terms of word sense ambiguity, approximately 261,642 lexemes in the HT dataset[2] have multiple possible senses, of which 32,679 have more than three senses. 70 lexemes are annotated with 100 or more senses each, with a word assigned to 345 possible senses. Although the average number of senses for each word of the whole HT dataset is slightly over 2.00, most commonly used words have multiple word senses.

In order to facilitate semantic analysis and make the category list more easily navigable for the users, a new set of 4033 thematic semantic categories was created for use in the HTST.[3] This is to create a ‘human-scale’ category set, i.e. one which comprises categories which a human would consider to be significant: neither too vague nor too detailed to be useful for general application. In practice, this involved drawing a notional ‘line’ through the HT, with categories above the line deemed cognitively significant to a human user, whilst the more intricate, often more technical categories below the line were (under the new set of headings) merged into their superordinate category above the line. Criteria used for this were based on the key concept of ‘human scale’ found in chapter sixteen of Fauconnier and Turner 2002. The resulting thematic hierarchy has five levels, which predominately map to the top five levels of the original HT hierarchy of semantic categories, and no subcategories. None of the lexis in the HT is lost, but rather some are less finely divided. Table 1 shows statistics of the distribution of semantic categories and their sub-categories of the thematic tagset for the top three hierarchical levels. Note that not all tags have five hierarchical levels. For example, tags *AB.02* (Birth) and *AB.08.d* (Development/growth/degeneration) have two and three levels of hierarchy respectively. This explains why the number of categories at layer three is reduced compared to that of layer two.

Users of the HTST can search using the thematic category set, as well as view results aggregated according to this set. The new categories are distinguished from the original HT categories using codes which combine letters and numbers, alternating between the two in successive levels of the hierarchy. For example, *Board game* is labeled BK.01.d.04.a as follows:

BK	*Leisure*
BK.01	*Entertainment*
BK.01.d	*A specific form of amusement/a pastime*
BK.01.d.04	*Game*
BK.01.d.04.a	*Board game*

The HTST tags text with both the full HT category and the thematic categories, so users can search on the most suitable level of detail for their needs. For example, HTST tags the word “mat” with the HT code “03.02.07.03.09.14-03" and corresponding thematic level code “AZ.06.f.05.m”. Considering the more practical

[2] The statistics is based on the HT dataset used in the current version of HTST.

[3] For definitions of the thematic categories, see: http://www.gla.ac.uk/media/media_405070_en.pdf and http://www.gla.ac.uk/media/media_405073_en.xlsx (both last accessed 19 April 2016).

Table 1
Distribution of semantic categories and their sub-categories of the thematic tagset for top three hierarchical levels.

Hierarchical layer	Number of cats	Min numb of sub-categories	Max numb of sub-categories	Average numb of sub-categories
layer-1	37	20	349	108.05
layer-2	331	1	124	10.48
layer-3	96	1	25	5.13

applicability of the thematic level categories for users, in this paper we evaluate the performance of the HTST based on these categories.

4. Architecture of the HTST system

The HTST tagger is based on a set of existing NLP tools developed at UCREL (Lancaster University). These are the Variant Detector (VARD) (Baron and Rayson, 2008), the Constituent Likelihood Automatic Word-tagging System (CLAWS) and USAS, which respectively provide functionalities of spelling normalization (particularly for historical text), tokenization and part-of-speech annotation, and semantic field annotation. These functionalities are required for pre-processing input text before we can apply the HT knowledge base for a deeper layer of semantic annotation. These tools introduce some errors in the pre-processing steps, which is inevitable for automatic tools, but in this paper we focus on the performance of the whole HTST system, and will not investigate the performance of the individual tools since this is reported in the relevant cited papers. Fig. 2 illustrates the pipeline architecture of the HTST annotation system.

Compared to other NLP toolkits, a unique component of HTST is the VARD software that is used for normalizing historical spelling variants of English. Spelling variation is a prevalent feature of historical varieties of most languages. This is especially the case for historical English, which has been shown to have high levels of spelling variation, decreasing over the Early Modern English period until around 1700 (Baron et al., 2009). VARD has been developed over a number of years to assist with the normalization of spelling in, particularly, historical texts. It utilizes methods commonly found in modern spellcheckers, such as phonetic matching, edit distance and letter replacement rules, but specialized for historical texts. The tool can also be trained to deal with spelling variation in different time periods and from different corpora, which improves the precision and recall of the automatic spelling normalization (Baron and Rayson, 2009; Hendrickx and Marquilhas, 2011; Lehto et al., 2010). In the HTST, VARD acts as

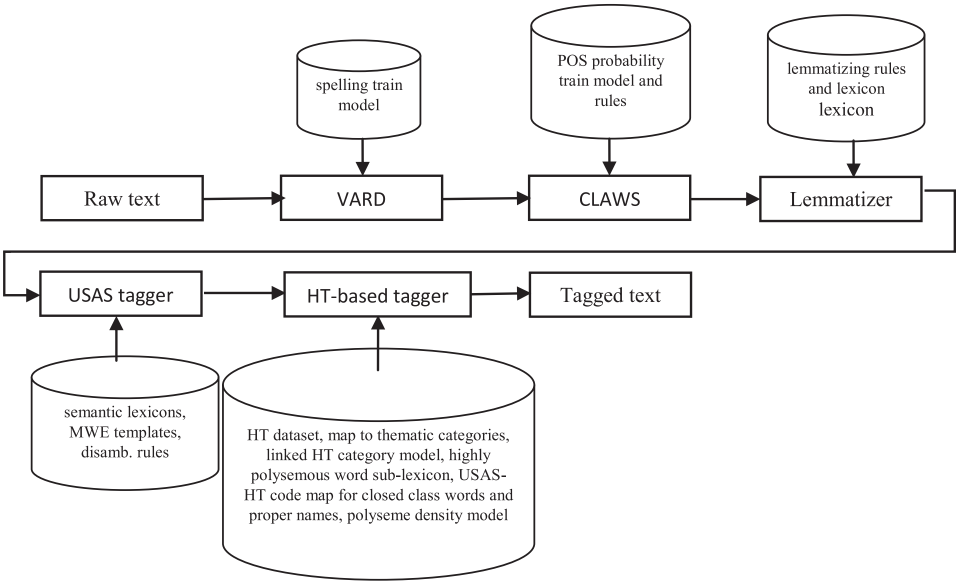

spelling train model
POS probability train model and rules
lemmatizing rules and lexicon lexicon
Raw text
VARD
CLAWS
Lemmatizer
USAS tagger
HT-based tagger
Tagged text
semantic lexicons, MWE templates, disamb. rules
HT dataset, map to thematic categories, linked HT category model, highly polysemous word sub-lexicon, USAS-HT code map for closed class words and proper names, polyseme density model

Fig. 2. Architecture of HTST system.

a pre-processor to other automated tools, which perform more accurately on normalized text versus the original text containing spelling variants. Previous studies have demonstrated that spelling normalization can significantly improve the performance of POS-tagging (Hendrickx and Marquilhas, 2011; Rayson et al., 2007) and semantic annotation (Archer et al., 2003).

One of the main components of the HTST is the USAS semantic tagger. Although it is now mainly used to semantically analyze general modern written texts and transcribed speech, USAS was originally designed for the content analysis of market research interview transcripts in order to bridge the gap between qualitative and quantitative survey methods. Rather than project specific categories, it applies a general purpose semantic taxonomy of 232 categories arranged in a hierarchical structure, with the top level containing 21 major domains (see Table 2) which in turn break down into three further levels of sub-division.

The USAS semantic taxonomy was originally based on Tom McArthur's *Longman Lexicon of Contemporary English* (McArthur, 1981) and allows the tool to distinguish between coarse-grained semantic fields rather than fine-grained word senses. For example, for the word 'bank', some printed dictionaries would distinguish between the conceptual categories of the financial institution and the physical high-street branches of the bank. The USAS tagger considers both of these senses as related to money and commerce (Archer et al., 2004).

The knowledge source of the system consists of a single word lexicon currently containing 56,318 items and a multiword expression (MWE) lexicon currently containing 18,971 templates, as shown below.

$$take * \_ * \{Np/P * /R*\}for\_IF\ granted\_*$$

Each single word or MWE has been manually assigned to one or more potential semantic fields (tags). MWEs are considered to be phrases or chunks, potentially discontinuous, which are assigned a single semantic tag. The MWEs mainly include phrasal verbs (e.g. 'stubbed out'), noun phrases (e.g. 'riding boots'), proper names (e.g. 'United States of America') and true non-compositional idioms (e.g. 'living the life of riley') (Piao et al., 2005).

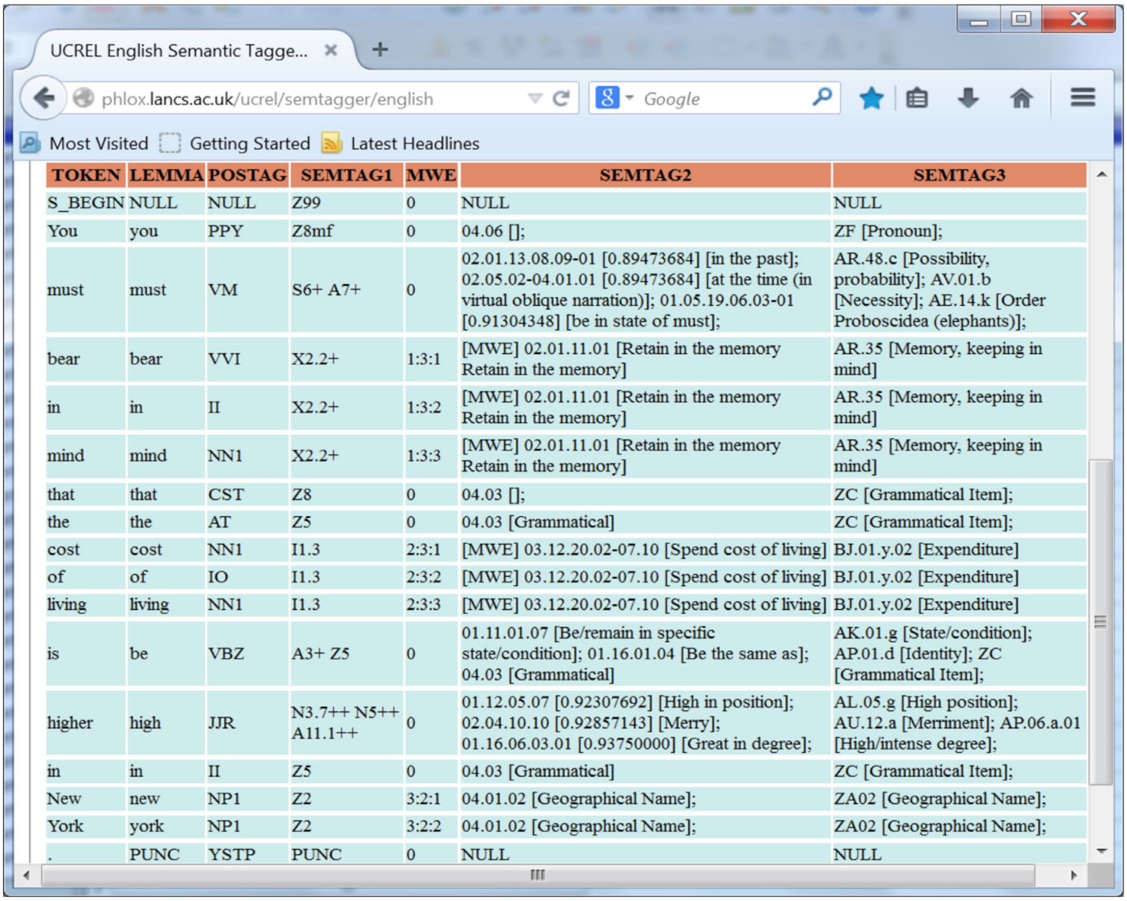
The USAS semantic tagger employs a combination of six methods to contextually disambiguate which of the potential tags is correct. A main method is the grammatical category of a word, hence we pre-process texts with the CLAWS POS tagger (Garside and Smith, 1997). For example, the word 'spring' can be partially semantically disambiguated if we know it is a verb versus a noun, to differentiate meanings such as movement-action (verb) versus metal-coil, season or water-source (all nouns). Subsequent disambiguation methods, applied in order, are:

(a) General likelihood ranking, derived from frequency information, past tagging experience and intuition;
(b) Overlapping template resolution, using length and span heuristics plus closeness of wildcard matching to prioritize which MWE template is selected;
(c) Domain of discourse, e.g. the topic of a text will help determine the relative ordering of tags
(d) Text-based disambiguation, using recurrence of an item in a text to assign the same semantic field in each case; and
(e) Manually created contextual rules, where immediate local context can determine the correct semantic field.

For a full description of the six methods with examples, see Rayson et al., (2004). Both CLAWS and USAS contain lemmatizers to enable improved dictionary look-up for grammatical and semantic tagging, although this is represented separately in our pipeline diagram.

Table 2
The USAS tagset top-level domains.

A: General and abstract terms	B: The body and the individual
C: Arts and crafts	E: Emotional actions, states and processes
F: Food and farming	G: Government and the public domain
H: Architecture, buildings, houses and the home	I: Money and commerce
K: Entertainment, sports and games	L: Life and living things
M: Movement, location, travel and transport	N: Numbers and measurement
O: Substances, materials, objects and equipment	P: Education
Q: Linguistic actions, states and processes	S: Social actions, states and processes
T: Time	W: The world and our environment
X: Psychological actions, states and processes	Y: Science and technology
Z: Names and grammatical words	


UCREL English Semantic Tagge...
phlox.lancs.ac.uk/ucrel/semtagger/english
Google
Most Visited
Getting Started
Latest Headlines

TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	NULL	NULL
You	you	PPY	Z8mf	0	04.06 [];	ZF [Pronoun];
must	must	VM	S6+ A7+	0	02.01.13.08.09-01 [0.89473684] [in the past]; 02.05.02-04.01.01 [0.89473684] [at the time (in virtual oblique narration)]; 01.05.19.06.03-01 [0.91304348] [be in state of must];	AR.48.c [Possibility, probability]; AV.01.b [Necessity]; AE.14.k [Order Proboscidea (elephants)];
bear	bear	VVI	X2.2+	1:3:1	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
in	in	II	X2.2+	1:3:2	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
mind	mind	NN1	X2.2+	1:3:3	[MWE] 02.01.11.01 [Retain in the memory Retain in the memory]	AR.35 [Memory, keeping in mind]
that	that	CST	Z8	0	04.03 [];	ZC [Grammatical Item];
the	the	AT	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
cost	cost	NN1	I1.3	2:3:1	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
of	of	IO	I1.3	2:3:2	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
living	living	NN1	I1.3	2:3:3	[MWE] 03.12.20.02-07.10 [Spend cost of living]	BJ.01.y.02 [Expenditure]
is	be	VBZ	A3+ Z5	0	01.11.01.07 [Be/remain in specific state/condition]; 01.16.01.04 [Be the same as]; 04.03 [Grammatical]	AK.01.g [State/condition]; AP.01.d [Identity]; ZC [Grammatical Item];
higher	high	JJR	N3.7++ N5++ A11.1++	0	01.12.05.07 [0.92307692] [High in position]; 02.04.10.10 [0.92857143] [Merry]; 01.16.06.03.01 [0.93750000] [Great in degree];	AL.05.g [High position]; AU.12.a [Merriment]; AP.06.a.01 [High/intense degree];
in	in	II	Z5	0	04.03 [Grammatical]	ZC [Grammatical Item];
New	new	NP1	Z2	3:2:1	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
York	york	NP1	Z2	3:2:2	04.01.02 [Geographical Name];	ZA02 [Geographical Name];
.	PUNC	YSTP	PUNC	0	NULL	NULL

Fig. 3. Sample output of the HTST annotation.

The existing NLP tools mentioned above form a basis upon which the HTST tagger has been developed. Currently, for a given input text, the HTST produces six layers of annotation, as shown in Fig. 3, where candidate HT tags (up to three) for each word are selected and sorted by likelihood scores (figures in the brackets). For example, for the word ‘children’ in the sentence “Mary has three children.”, HTST produces the following information:

1) Lemma ‘child’
2) Part-of-speech ‘NN2’
3) USAS semantic tag ‘S2mf/T3-S4mf’
4) Multiword expression flag ‘0’
5) HT sense code ‘01.04.04.04’
6) Thematic level sense code ‘AD.03.d’

5. Disambiguation of HT semantic categories for words

In the HTST tagger, we have implemented a number of word sense disambiguation methods. In detail, the HTST combines the following disambiguation methods:

1) Employ manually crafted sub-lexicons of 200 words which provide core HT categories of highly polysemous words, such as ‘come’, ‘make’, ‘take’ etc.

2) Map USAS tags to HT categories for closed-class words, where stable relations exist between these two sets of tags. This method is mostly applied to function words and proper nouns.
3) Polysemy density model which provides most likely HT category code out of the HT's top three layers.
4) Context-based disambiguation model based on statistical distance between the HT semantic categories, which function as brief definitions of the categories, and the context words.
5) Context-based disambiguation statistical model based on a HT-USAS semantic tag association model, which is extracted from the OED word sense definitions.
6) Time-filtering of the HT categories, which aims to remove irrelevant categories of word senses with respect to time of word usage, e.g. archaic meaning for contemporary word usage or newly coined meaning for historical usage of words.
7) Mapping the full HT categories into the thematic categories, which helps to disambiguate word senses by eliminating extremely fine-grained semantic distinction.

The first main disambiguation approach is to use a manually compiled sub-lexicon and tag mapping to deal with highly polysemous words (methods 1 and 2 on the above list). Such words can have a maximum of more than a hundred HT meanings assigned to them. These include "*set"* which has 425 possible meanings in the *Historical Thesaurus*, and "*run"*, which has 337 possible meanings. The level of detail achieved by the *Historical Thesaurus* in these categorizations is essential for accurately representing the complex semantics of the language, and predominately follows the *OED*'s equally detailed division of word senses. However, the exceptionally fine-grained nature of this categorization scheme can lead to difficulty in disambiguating word senses using automatic algorithms (and, indeed, human experts may disagree on precise categorization of given instances). Along with the polysemy density model (method 3), the sub-lexicon and tag mapping provide an effective method for disambiguating the word sense for highly polysemous words.

HT head words are the words used to define categories, and were composed by the editors of the *Thesaurus* during its creation. The appearance of a word in a category heading does not preclude the inclusion of the word within the category. For example, category 03.01.01.03.10 'ancestor' includes the word *ancestor.* The use of a word in a category heading is not a reliable indicator that the category contains the most common or important sense of that word, but has been trialled in this system as a clue and indicator (technically as part of the feature set) for selecting senses of polysemous words to be more highly weighted. The contextual disambiguation method based on the head words (method 4) works as follows:

1) For a given word and for each of its candidate HT categories: Extract all possible parent HT categories and collect their headings (key words that define a HT category), including the headings of the HT category under consideration. Words of the headings form a feature set $\boldsymbol{HW_i}=\{h_1, h_2, \ldots, h_m\}$.
2) Collect up to five content words from each side of the key word/MWE. Together with the target word/MWE w_t, they form a context feature set $\boldsymbol{CW}=\{w_t, w_1, w_2, \ldots, w_n\}$.
3) Measure Jaccard Distance (Levandowsky and Winter, 1971) between $\boldsymbol{CW}$ and each $\boldsymbol{HW_i}$, and select the candidate categories (up to three) that have the closest distances to the context. Jaccard Distance is calculated as below:

$$Jaccard\_Distance_i = (|CW \cup HW_i| - |CW \cap HW_i|)/(|CW \cup HW_i|).$$

A key disambiguation method is based on the OED data (method 5), which is made possible due to the HT category information contained in OED entries (for further details about the link between HT and OED, see website: http://public.oed.com/historical-thesaurus-of-the-oed). Most OED headwords are linked to one or more HT categories, e.g. "ancestor" is linked to HT code *03.01.01.03.10* which denotes the concept "Kinship/relationship->Ancestor". Our approach is first to tag the word sense definitions in the OED using the USAS tagger, then extract the statistical association metric between the HT categories of the headwords and the USAS tags contained in the definition entries. Below is a USAS-tagged OED definition of the word "ancestor", which illustrates a strong statistical association between the HT code *03.01.01.03.10* and USAS category of *S4* (kinship).

ancestor_03.01.01.03.10_n:

One_Z8 from_Z5 whom_Z8 a_Z5 person_S2mfc is_Z5 descended_M1,_PUNC either_Z5 by_Z5 the_Z5 father_**S4m** or_Z5 mother_**S4f**;_PUNC a_Z5 progenitor_**S4**,_PUNC a_Z5 forefather_**S4m** ._PUNC (_PUNC Usually_A6.2+said_Q2.1 of_Z5 those_Z8 more_N5++ remote_N3.3+than_Z5 a_Z5 grandfather_**S4m** ._PUNC)

_PUNC Also_N5++,_PUNC of_Z5 animals_L2mfn,_PUNC and_Z5 fig._Z99 as_Z5 spiritual_S9 ancestor_**S4/**T1.1.1 ._PUNC

As shown in the sample, the *S4* tag occurs six times in the definition, implying a strong association with the HT concept of ancestor (*03.01.01.03.10*). Note that the *Z5* tag indicates grammatical function words, therefore they are excluded from the association extraction process. Table 3 shows a sample from a log-likelihood association table between HT categories and USAS tags extracted from the entire OED definition data, where the codes denote semantic meanings as shown below.

HT Codes (same sequence as in table):

01.01.11.02.07: Wind
01.11.02.02: Destruction
01.09.10.02: Freedom from impurities
01.15.18.01: Calamity/misfortune
01.02.02: Biology
01.09.10.03: Pest control
02.05.05.04: Obstinacy/stubbornness
01.16.06.09: Decrease/reduction in quantity/amount/degree
01.15.22.01: Skill/skillfulness
02.06.06: Poverty

USAS Codes (same sequence as in table):

W4: Weather
A1.1.2: Damaging and destroying
B5: Clothes and personal belongings
A1.4-: Chance, luck (negative)
L1-: Life and living things (negative)
A2.1-/X2.1: Affect: Modify, change (negative) / Thought, belief
N5-/A2.1: Quantities (negative) / Affect: Modify, change
X9.1 +: Ability:- Ability, intelligence (positive)
I1.1-: Money: Affluence (negative)

In the HTST, this association table is used to find the candidate HT category which has the greatest mean association score with the surrounding USAS tags within its context of sentence.

The novel time filtering approach (method 6 in our list of disambiguation methods employed by the HTST) is possible because the HT dataset contains records of the periods of time during which a given word sense appears. Our approach is to restrict the range of candidate word senses to those which appear within a certain time-window around the publication time of a given text. For example, if a text was published in 1810, then we set a time range of 1750 to

Table 3
Sample of association between HT categories and USAS tags.

Co-occurring tag pair	Co-occur. freq	HT code freq	USAS code freq	Log-likelihood
01.01.11.02.07_W4	387	466	4815	2435.92859
01.11.02.02_A1.1.2	414	727	8489	1925.78816
01.09.10.02_B5	502	626	17490	1895.23302
01.15.18.01_A1.4-	180	301	422	1862.89737
01.02.02_L1-	289	336	5470	1768.35265
01.09.10.03_B5	455	555	17490	1732.90144
02.05.05.04_A2.1-/X2.1	173	459	354	1712.51737
01.16.06.09_N5-/A2.1	256	360	3632	1698.58208
01.15.22.01_X9.1 +	310	394	7466	1670.79259
02.06.06_I1.1-	199	349	1240	1661.24810

1850 and filter out the candidate word senses which were not used within this period. Our assumption is that such a method would reduce some of the noise during the automatic word sense disambiguation process.

Finally, the full HT categories selected by the HTST are mapped to the thematic HT categories (method 7), which provides a more practical granularity of semantic category classification and which helps to disambiguate word senses. In the evaluation in the next section, we use the thematic HT categories to test the accuracy of the HTST tagger's output.

6. Evaluation

In this section, we describe our evaluation of the HTST, including test data preparation and evaluation criteria (in Section 6.1), statistical results of the HTST performance and the impacts of the main disambiguation methods implemented in the HTST (Section 6.2), and software design to improve the runtime speed of the HTST software (Section 6.3).

6.1. Test data preparation

For the evaluation of the performance of the HTST, we selected ten texts from different genres/domains and publication times as test data to be manually annotated. The test texts were restricted to approximately 1000 words in length to make them manageable for the manual annotators; where texts were longer, excerpts were selected. Texts ended at a sentence (and often a paragraph) boundary so that there were no unintended sentence fragments. In order to ensure the objectivity of our evaluation, these test data were not used in the development of the HTST system.

As part of the SAMUELS project, in which the HTST was developed, three sub-projects conducted research using the semantically-tagged Hansard Corpus (Alexander and Davies 2015; available at http://www.hansard-corpus.org; last accessed 19 April 2016) covering the years 1803 to 2005, and so text selection was guided by the requirement to evaluate the tagger on a variety of writing styles across the period covered by this corpus. Eight test texts were selected based on this criterion, which were published from the 19th century onward, and we consider these as contemporary English for the purposes of our study (because spelling is standardized in English by this time).

In addition to the test texts of contemporary English, two Early Modern English (EModE) test texts were selected: a news text dated 1621 (sourced from the Burney Collection database) and a comedy drama dated 1623 (sourced from Early English Books Online (EEBO)). These genres were chosen because they presented the HTST with different levels of challenge in semantic disambiguation. The news text is written in continuous prose and purports to be a mainly factual report of events which have taken place (in our sample, a fire in the city of Paris). It contains relatively little figurative language compared to the comedy drama, which is more densely packed with metaphor and other figurative devices.

In order to help evaluate the correct meanings in context, the researchers made use of the substantial body of digitized texts from the Early Modern period on EEBO (to examine other cases in context of unfamiliar words in the sample texts), for both EModE samples. Additionally, for the comedy drama sample, which was from a play by William Shakespeare, the corpus-based glossary of Shakespearean meanings by Crystal and Crystal (2002) and the notes from a scholarly edition of Shakespeare's plays (Greenblatt et al. 1997) were also used to help determine correct meanings. Table 4 lists the ten test texts.

In order to check the accuracy of automatic annotation of the HTST tagger, the test texts were manually annotated with the HT semantic categories. Table 5 shows a sample of manually annotated text. Where the candidate code suggested by the POS was NULL, as in lines 2, 3, 6, 7 and 8 in Table 4, the HTST does not identify the lexical item sufficiently well to suggest a candidate meaning code. NULL cases in our example are mostly metadata which help zone the text, e.g. the tags S_BEGIN and S_END, which are not of interest for analysis in any case, and required no action from the researchers. The lexical item 'Newes', in line 2, would, however, be of interest. The spelling was not normalized to 'News' during the VARD stage of the HT tagging process (discussed above in Section 4 and shown in Fig. 1), which would probably have allowed the HTST to suggest a correct candidate HT code. In this case the correct candidate codes were manually added in the Correct-Tag column by the researcher after scrutinizing the possibilities in the online HT. In this case the most appropriate codes for the context were 03.09.05.09 [News/tidings] or

Table 4
List of test texts used for HTST evaluation.

Test text	Genre	Description
File 1	Biography	By Stratemeyer, Edward. 1904. *American Boy's Life of Theodore Roosevelt.* Boston, MA: Lee and Shepard [From the Project Gutenberg edition].
File 2	Fiction (General)	By Dickens, Charles. 1852. 'I In Chancery'. From *Bleak House*, [Project Gutenberg edition].
File 3	Fiction (Humor)	By Wodehouse, PG. 1915. Chapter 1. From *Something Fresh*, [Project Gutenberg edition].
File 4	Political speech	By Scott, William. 1820. Choice of a Speaker. HC Deb 21 April 1820, volume 1, columns 2−9.
File 5	Political speech	By Blair, Tony. 2010. International Terrorism and Attacks in the USA. HC Deb 14 September 2001, volume 372, columns 604−16.
File 6	Historical writing	By Gibbon, Edward (revised with notes by H.H. Milman). 1845 [1782]. 'II.II The Internal Prosperity In The Age Of The Antonines'. From *The History of The Decline and Fall of the Roman Empire*, volume 1. [Project Gutenberg edition].
File 7	Journalism	By Silverstein, Ken. 1998. 'The Radioactive Boy Scout'. *Harper's Magazine*, November 1998.
File 8	Journalism	By Moran, Caitlin. 2010. 'Drinking is like a mini-break—as exhilarating as spending three days sightseeing in Rome'. *The Times* (Opinion column), 17 July 2010.
File 9	News	By author unknown. 1621. 'Newes from France'. The Burney Collection of English Newspapers, British Library.
File 10	Comedy drama	By Shakespeare, William. 1623. *The Merry Wives of Windsor*, II:ii [EEBO First Folio edition[a]].

[a] The play was first produced in 1597−98, and there is a quarto edition dated 1602 (see the Database of Early English Playbooks at http://deep.sas.upenn.edu/index.html, last accessed 6 July 2016).

03.09.05.09-01 [piece of]. Either would be correct in the context, and, although one or the other would suffice, to enable the HTST to 'learn' most effectively both were added manually. The selection of correct candidate codes in cases such as this can be subjective and open to debate, as we cannot be sure of either the author's intended meaning (s), or the way in which the meaning(s) would be inferred by a contemporaneous audience. These issues could themselves be the subject of lengthy discussions. A practical way forward was achieved through the application of inter-rater agreement (discussed further below).

HT codes beginning '04' are not part of the taxonomy in the HT itself (Kay et al. 2016), but are a separate set designed for the SAMUELS project to accommodate frequently-occurring lexical items that have little or no semantic content or to group together unmatched lexical items in a convenient way. Many grammatical words are subsumed under the category 04.03, including prepositions (e.g. 'from', line 3), conjunctions (e.g. 'and', line 8) and articles (e.g. 'the', line 9). Periphrastic 'do' is also assigned to the 04.03 Grammatical category, as in "The Change Bridge did no less feel the force". Geographical names are subsumed under a single category, 04.01.02 (e.g. 'France', line 4).

The final lexical item in our example, 'Elements' (line 10), has three candidate HT codes suggested by the HT: 01.01.10.02.02.02 [1.00000000] [Sphere of ancient astronomy]; 01.10.01.01 [1.00000000] [Alchemical elements]; and 01.10.02.17-01 [1.00000000] [elements]. The number in brackets after the semantic category label (which is always between 0 and 1) indicates the likelihood of the code being correct. The lower the number, the more confidence there is that the meaning in context is correct. In this case the confidence of the tagger is equal for all three candidate codes. These were checked manually by the researcher who determined that the first two codes were indeed possible meanings in the context (either would be correct, and one is not obviously more correct than the other—though we cannot discount the possibility that contemporaneous audiences would disagree). The third suggested candidate meaning code relates to elements and compounds but has a meaning that does not start until 1724 (according to the OED data imported into the HT), more than 100 years later than our sample, so it was discounted as a possible candidate meaning. The two correct candidate meaning codes were then entered in the Correct-Tag column.

Table 5
Sample of manual annotation of HT semantic category codes.

Line no.	Token	POS tag	Manually assigned semantic tag
1	S_BEGIN	NULL	
2	Newes	NP1	03.09.05.09 [News/tidings]; 03.09.05.09-01 [piece of]
3	from	II	04.03 [Grammatical]
4	France	NP1	04.01.02 [Geographical Name]
5	0	YSTP	
6	S_END	NULL	
7	S_BEGIN	NULL	
8	and	CC	04.03 [Grammatical]
9	the	AT	04.03 [Grammatical]
10	Elements	NN2	01.01.10.02.02.02 [Sphere of ancient astronomy]; 01.10.01.01 [Alchemical elements]

Metaphors in the sample texts presented both a challenge (both to the HTST and to the researchers) and an opportunity to gain insight into the pervasive use of metaphor in English. In some cases, the HTST suggested candidate meaning codes corresponding to the metaphorical meaning, along with others corresponding to the literal meaning. In such cases, the researchers entered just the metaphorical meaning code in the Correct-Tag column, not the literal meaning, as the metaphorical meaning is the one that is correct in the context. Where literal meaning codes only were suggested for metaphorical lexical items, the researchers checked whether or not a meaning code for metaphorical use exists in the HT, and if so it was entered in the Correct-Tag column. If no metaphorical meaning code exists in the HT, it is beyond the capabilities of the HTST to suggest one, and therefore the literal meaning code most closely corresponding to the source of the metaphor was entered in the Correct-Tag column (with the additional cognitive step of interpreting the metaphorical meaning left for the reader/audience to make). Work on this was informed by a separate AHRC-funded project using the HT, *Mapping Metaphor with the Historical Thesaurus* (Anderson et al., 2015), whose findings demonstrate the complex interrelationships of metaphor in both the HT database and the history of English. Although the version of the HT database employed for building the HTST does not contain information on the metaphorical links between categories, it would be possible to include this information in the instance of further development of the system. This should lead to a better method of automatically identifying cases in which competing semantic tags are the result of metaphorical word use.

Each text sample was independently checked by two researchers, and their decisions on the evaluation of the annotations compared afterward. Tricky cases were discussed (for example, in cases where the nuances of a particular semantic meaning were unclear from the context, and where more than one HT category could reasonably be considered equally correct). The researchers carrying out the evaluation had expertise in historical English and experience in working with older texts, and in nearly all cases there was inter-rater agreement over the semantic category which was most correct in the given context. For a minority of cases the most correct meaning remained debatable in the context, and we then applied the HT category which would be the least contentious, whilst acknowledging that another meaning could apply but the evidence for its application was insufficient. Finally, if different annotators chose different HT codes for the same words but did not contest the other's choice, their annotations were merged, assuming both of them are correct.

It is a highly challenging and time-consuming task to produce manually annotated test corpus data due to the highly fine-grained nature of the HT semantic classification scheme. Therefore, the size of the manually analyzed test data is limited. However, the high quality of manual annotation carried out by the linguistic experts and the wide coverage of the data both in terms of genre/domain and publication time enabled us to test the HTST's performance on different types of text using relatively small amounts of test data. (The manually annotated test data is available at URL: https://github.com/UCREL/SAMUELS).

As discussed above, the HT database can contain multiple correct semantic codes for a single word in a given context, all of which provide correct word senses or components of word sense in the context of its use. The nature of any natural language is that meaning is not precise and categorical, but is emergent in the sense of any complex system. During the manual annotation, therefore, the annotators sometimes assigned different HT codes for a word from different branches of the hierarchical HT semantic classification structure, all of which provide correct senses of the word. For example, the two annotators assigned HT code pairs of '03.09.05.04 [Report information], 03.09.11-01 [Journalism]' and '02.07.03.03 [Narration], 03.09.05.04' to the word 'report'. Merging together, this word has three HT codes assigned: 03.09.05.04, 03.09.11-01 and 02.07.03.03. Here, the first two codes and the third one derive from different top-level semantic categories explained in Section 3: 'The Social World' (with 03 code prefixes) and 'The Mental World' (02 code prefixes) respectively. There are numerous such cases, as shown by examples in Table 6, where HT tags have the following definitions.

Table 6
Sample of diverse of HT codes assigned to words during manual annotation.

Word	HT codes	Description
Accident	01.15.18.01-02.10.05, 01.11.04-02, 01.15.18.01-02.10.05	Same top category, but from different second level branches
Plainly	01.09.08.11-10, 02.01.07-01, 03.09.02.01	All codes from different top categories
Surprised	02.01.14.02-05, 02.01.14.02-04, 02.01.14.02	Share main code, but different sub-categories
Astonishment	02.01.14.08, 01.03.02.01.02-04, 01.03.03.04.21-01, 02.04.21.08-02	Codes from wide range of locations in HT sense structure
Take	01.11.03.02-03, 01.15.06-03, 02.06.08-12	Highly polysemous word can have codes from diverse HT sense classification branches

01.15.18.01-02.10.05: Calamity/misfortune: a mishap/unlucky accident
01.11.04-02: Occurrence: an occurrence/event
01.09.08.11-10: Visibility: sightworthiness
02.01.07-01: Perception/cognition: theory of
03.09.02.01: Manifestness
02.01.14.02-05: Surprise, unexpectedness: one who surprises
02.01.14.02-04: Surprise, unexpectedness: feeling of surprise
02.01.14.02: Surprise, unexpectedness
02.01.14.08: Feeling of wonder, astonishment
01.03.02.01.02-04: Degree/type of madness: mental prostration/paralysis
01.03.03.04.21-01: Anaesthetization/painkilling etc.: anaesthetization
02.04.21.08-02: Dismay: consternation
01.11.03.02-03: Source/origin: source of material thing
01.15.06-03: Carrying out: of justice, etc.
02.06.08-12: Acquisition: that which is obtained/acquired

When we merged manual annotations of the same test text carried out by different annotators, we found that many content words (nouns, verbs, adjectives and adverbs) were annotated with more than one HT code, up to six different codes for single words.

Such a diversity of HT codes sharing similar senses that are applicable to a given word presents a challenge for our evaluation. An implication of the HT code diversity is that, in many cases, there exist multiple correct HT codes for a single word. The question is: do we request that the annotation tool find all of the correct HT codes for each word? Obviously, it is impractical to achieve this with an automatic tool, as it is very difficult even for human linguistic experts. During the manual annotation, very often the individual annotators found that the possible codes were present in closely related categories within the HT, and thus that the higher-level components of the category codes were the same for multiple active fine-grained meaning distinctions. This means that possible senses of any given word often shared digits at the beginning of their codes, as in the case of ‘surprised’ in Table 6.

A more practical approach would be to require the semantic annotation tool to find the whole or part of the correct HT code set for each word, and we adopted this approach in our evaluation. Considering the multiple possible correct HT codes for a given word, if the annotation tool assigns at least a correct HT semantic code within the top three most likely candidate tags of a word (ranked by statistical likelihood scores), we considered it to be successful. For cases where more than one correct HT code is identified by the tool, it is still considered to be a single successful case. Although the high level of HT granularity creates difficulty with assigning a single ‘correct’ tag to many words, this granularity is still valuable in the weighting of senses using polyseme density (Section 5, method 3 above).

6.2. Impacts of disambiguation methods

In our evaluation, we focused on examining the impact of four main disambiguation methods on the performance of the HTST. They are:

1) The sub-lexicons, mapping list between some USAS tags and HT categories, and polysemy density list, including methods 1), 2) and 3) described in Section 5.
2) Context-based disambiguation model based on the HT head words and the context words.
3) Context-based disambiguation model based on the HT-USAS semantic tag association model, which is extracted from the OED word sense definitions.
4) Time-filtering of the candidate HT categories.

In order to assess the impact of the word sense disambiguation methods, we first estimated a baseline performance of the HTST by randomly selecting HT categories from the dataset for given words, i.e. the accuracy of the annotation without using any of the disambiguation methods. As a result, the HTST obtained an average baseline accuracy

Table 7
Impact of resources and disambiguation methods in comparison with baseline performance.

Test file	File 1	File 2	File 3	File 4	File 5	File 6	File 7	File 8	File 9	File 10	Avg.
Words	1027	932	1004	1027	1032	1567	1005	877	1020	1225	
Baseline (%)	12.27	19.42	14.44	14.70	16.57	14.42	16.72	12.88	20.00	12.65	15.41
Sub-lexicons, polysemy density, tag mapping (%)	66.50	68.56	68.03	72.15	70.83	71.35	65.07	67.16	78.63	68.16	69.39
Context/headwords based disambig. (%)	76.34	77.25	79.38	84.03	78.20	81.49	74.23	74.80	90.20	81.80	79.77
OED training (%)	76.14	77.79	79.28	84.32	78.00	82.32	74.23	74.91	89.80	81.71	79.85

of 15.41% on the ten test texts. We then added the above-mentioned four main methods one by one and re-calculated the accuracies in each step. Table 7 lists the results (refer to Table 3 for the contents of the test files) showing the impacts.

As shown in Table 7, the HTST achieved average accuracies of 69.39%, 79.77% and 79.85% after adding the disambiguation measures numbered 1), 2) and 3) respectively. This result demonstrates a significant improvement for the sub-lexicons etc. of method 1) over the baseline as well as a noticeable improvement resulting from the context based disambiguation method 2). The results of method 3) are more complicated. Overall, this method only marginally improved the average accuracy, from 79.77% to 79.85%. But a closer observation reveals that this method has a greater impact on contemporary texts than on historical texts. If we only consider the eight contemporary test files (from file 1 to file 8), the average accuracy is improved from 78.22% to 78.37%. On the other hand, this method has a slightly negative impact on the historical data (see file 9 and file 10). There are two factors accounting for this result. The first one is that the USAS semantic tagger performs less accurately on historical data, thus affecting the performance of the OED data based disambiguation, which is based on the USAS tags. The second one is that, because the OED word sense definitions are written in contemporary English, as training data they therefore do not reflect the features of historical English texts published hundreds of years ago. This shows that our current USAS tags based association model is suitable for processing contemporary text, but not suitable for tagging historical data.

Next, we tested the impact of the time-filtering method. In detail, we tagged each of the ten test files by changing a pair of time-boundary variables, namely lower and upper time boundaries, around the publication time of the texts. For example, for a test text published in 1621, the upper time boundary increases by 50 years each step from 1650 until the upper limit of 2000 is reached (the date considered to be ‘the present’ in the HT dataset). On the other hand, the lower time boundary reduces by 50 years each step from 1600. In this way, we obtain the HTST’s accuracy for all possible time ranges within the search space, as shown in Figs. 4 and 5 (the lower time boundaries earlier than 1000 did not affect the accuracy, so they are not shown in the figures). The tables in these two figures reflect the impact of the time filtering algorithm on the historical text (published in 1621) and contemporary English text (published in 1820). The bottom-left corners of the figures represent the narrowest ranges between the upper and lower time boundaries, and the lighter colors indicate the higher accuracies. From Figs. 4 and 5, we can see that the accuracy peaks at certain time-ranges for both test files: 91.08% in Fig. 4 and 86.37% in Fig. 5.

We repeated the same experiment for every test file, and found the accuracy peaks at certain time-ranges on every file, improving the average accuracy by 1.72% on average from 79.89% to 81.61%, as shown in Table 8. This indicates the time filtering approach can be effective for all types of texts. Table 8 also shows the HTST’s tagging accuracies obtained by using the time filtering method for each of the test files in comparison with those obtained without it. As the results show, proper time filtering can improve the HTST’s performance substantially on different types of texts, both contemporary and historical. For example, the time filtering improved the accuracy of File 2 (general fiction dated 1852) by 3.54%.

A technical challenge in applying the time filtering disambiguation method is automatically detecting the publication time of a given document. For well-designed corpora, such information can be encoded in the data, but in other situations, techniques for determining publication time of documents would be necessary for applying the time filtering method.

Fig. 6 summarizes the improvements of performance of the HTST tagger achieved by combining more and more disambiguation methods in the sequence of a) sub-lexicon, polysemy density and tag mapping, b) context-words based disambiguation, c) USAS-HT tag collocation in OED sense definitions, and d) time filtering of word sense usage in HT dataset, which correspond to the green, dark-blue, light-blue and yellow lines respectively. When

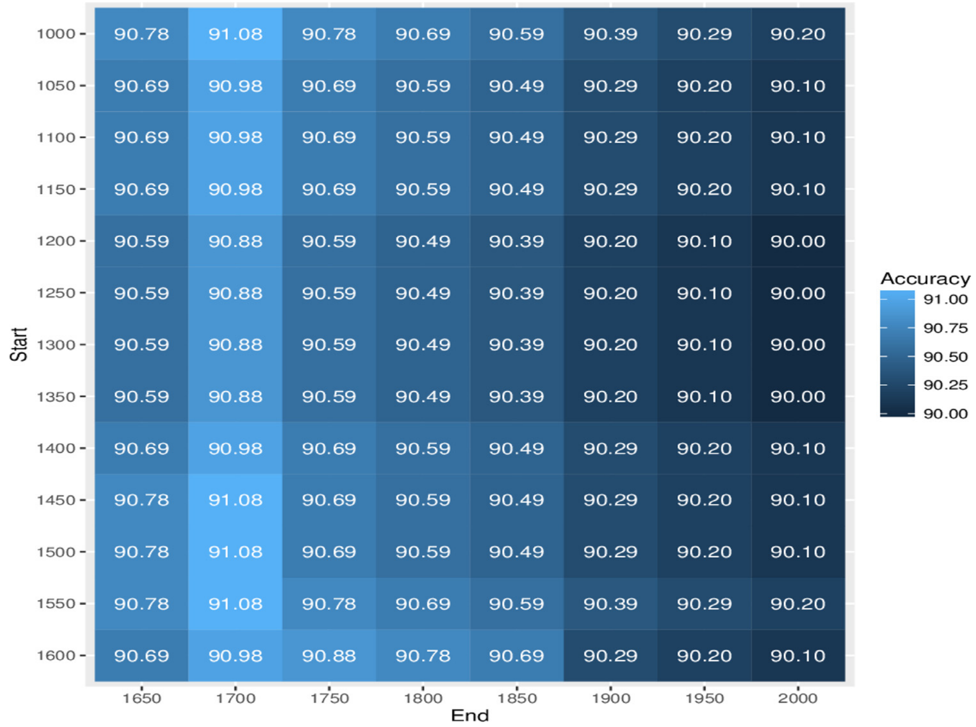


Fig. 4. Search for optimal performance in time range space on historical data (File 9: news published in 1621).

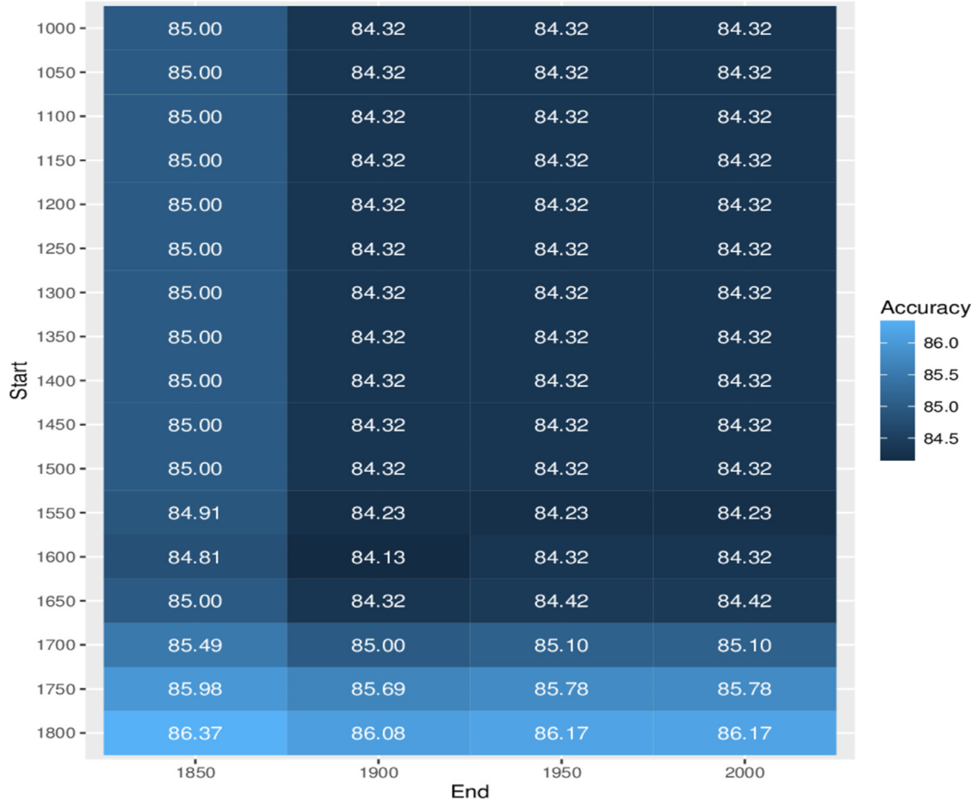


Fig. 5. Time range search space used for searching optimal performance (File 4: Parliament speech in 1820).

Table 8
Improvement of performance achieved by time filters.

Test file	File1	File2	File3	File4	File5	File6	File7	File8	File9	File10	Average
Pub. date	1904	1852	1915	1820	2001	1845	1998	2010	1621	1623	
Accuracy with no time filtering	76.14	77.79	79.28	84.32	78.00	82.32	74.23	74.91	90.10	81.80	79.89
Accuracy with time filtering & time range	(1900–2000) 77.12	(1750–1860) 81.33	(1850–2000) 80.68	(1800–1850) 86.37	(1850–2000) 80.68	(1800–1850) 84.88	(1950–2000) 75.42	(1900–2000) 76.62	(1450–1700) 91.08	(1500–1850) 81.96	81.61

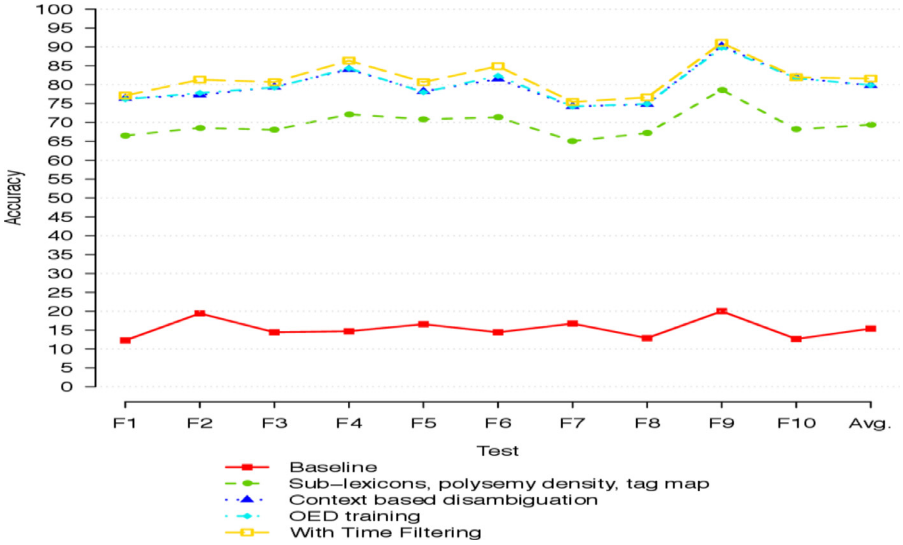


Fig. 6. Improvements of four disambiguation algorithms in comparison with baseline.

combined together, these disambiguation methods have improved the tagger's performance substantially on average, from 15.41% of the baseline to 81.61%: an improvement of 66.20 percentage points (or approx 400%).

Finally, we tested the significance of different methods against the baseline performance, using Chi-squared tests with Bonferroni correction to compensate for the use of multiple hypotheses. Fig. 7 shows the result, where

Method_1: Sub-lexicons, polysemy density, tag mapping.
Method_2: Context/headwords based disambiguation.

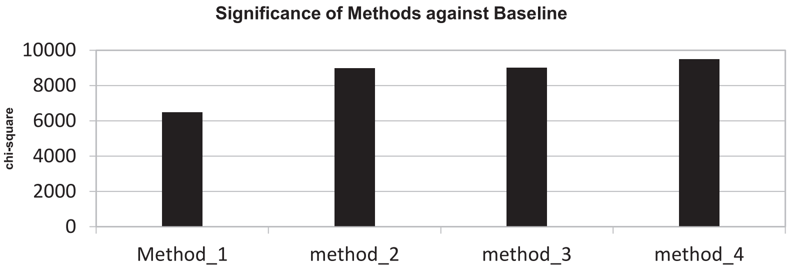


Fig. 7. Significance test of different methods against the baseline performance using chi-square metric.

Method_3: OED word sense definition data training.
Method_4: Time filtering.

As Fig. 7 illustrates, all refinements are shown to be significantly higher accuracy at the 1% level across our test corpus (one-tailed; $X > 6490$; $p < 0.000001$).

Despite the encouraging performance of the current version of the tagger, it needs further improvements in order to provide a reliable annotation system for a wider range of practical applications. For example, when only the first tag of the automatically produced candidate tags is considered, the average accuracy drops to 66.82%. More efficient algorithms and methods are needed to rank more correct senses to the top in the candidate list. In light of our experience, cleaner training data and more efficient context-based disambiguation algorithms would be needed to achieve a higher accuracy of the HTST's performance.

6.3. Overview of main error types

Our manual analysis of the errors in the annotation of the HTST reveals the types of challenge faced by semantic annotation tools, some of which are unique to this system. These can be exemplified using one of the fiction test texts, Dickens' *Bleak House*. As previously mentioned, metaphor and metonymy are the most common causes of error in the annotation. For example, for the word *face* in the phrase 'the face of the earth', the tagger weights most highly the category 'Face, facial expression' (AB.17.e.01.a). In this instance, the source domain for the metaphor has been identified, and is arguably correct. However, the human taggers prefer the metaphorical extension of *face* found in the category 'Surface' (AL.05.c.01), which the HTST returns as the third candidate tag in this instance. Metaphorical interference also appears to be present in the mislabeling of *broke* in 'if this day ever broke', for which the HTST returns 'Breaking/cracking' (AK.02.b.01) as the most likely sense.

Difficulty with metonymy can also be found with the word *soot*, for which the tagger most highly weights 'Black/blackness' (AJ.09.e.02) followed by 'Granular texture' (AJ.04.c), identifying senses of *soot* for which a quality of the substance is expressed using the term for the substance itself. The correct tag, 'Burning, fire, flame, ash' (AJ.03.c.02.c), is given in third position. One of the most likely contributing factors to this error are that senses of soot are found close together in the thematic level semantic hierarchy (all three are found in the AJ 'Matter' category). Also contributing to the higher weighting of 'Black/blackness' is the presence of the word *black* itself earlier in the sentence: 'Smoke lowering down from chimney-pots, making a soft black drizzle, with flakes of soot in it[...]'. Similar confusion between metonymic senses can be found in *river*, for which the first given tag is 'Action/process of flowing' (AJ.05.h), and the correct tag, 'Rivers, streams' (AA.04.b), is returned as the second likely candidate. Here, as before, the phenomenon (river) is used to represent one of its most salient properties (flowing motion), while the more literal tag is the one agreed as most correct by human taggers.

Some errors arise from the difficulties in identifying and labeling multi-word expressions (MWE). These come in two main forms: identification of an MWE where one is not present; and parsing and individual labeling of words which could be conceived as forming a set phrase. The HTST is intended to treat words individually wherever possible. In some instances, however, it is debatable whether individual words should have been grouped together. The words in 'Lincoln's Inn Hall' are tagged separately with varying degrees of accuracy (the first tags for its component words are 'Geographical Name' (ZA.02), 'Public lodging-place' (AZ.06.e.01.a), and 'University administration' (BE.05.b) respectively). It is arguably the case that these could be labeled as a single MWE with the 'Geographical Name' tag to indicate its status as a location. The opposite has occurred where 'in the streets' has been erroneously labeled as an MWE related to a way of dealing stocks and shares (BJ.01.aa) when individual tagging of each word would have been more appropriate.

These types of error are applicable to more than just the HTST. However, there are a couple of errors which are particular to the HTST pipeline. Where words are highly polysemous, the tagger can refer to a list of 'most likely' senses (see Section 5, method 1), and may apply one of these where it is actually not appropriate. This is the case for the adverbial *over* in 'Michaelmas term lately over', which is tagged with the sense 'Movement over/across/through/past' (AN.05.i), 'Frequency' (AM.10), and 'Transference' (AN.06) from the most likely sense list. Also present in 'Michaelmas term lately over' is a second HTST-specific error, in which the word *term* is labeled with 'Words and phrases' (AX.19). This appears to result from the headword measures (see Section 5, method 4), for which a tag has been weighted more highly because the word to be tagged (i.e. *term*) itself appears in the category or subcategory heading (‘term/expression’, 02.07.04.05.02-05). It is often difficult to tell where this has occurred since it is frequently true that the search term does appear in the correct HT heading, as is the case with *implacable* (’implacable’, 02.05.05.04-04), *weather* (’Weather’, 01.01.11.02), and *mud* (’Mud’, 01.01.07.04.06.03), all found within the first thirty words of this passage. Future development of the HTST will seek to fine-tune the tagger’s judgment of how highly to weight measures such as this and the use of the polysemous words list.

6.4. Issue of speed as a resource-intensive software

As detailed in Section 4, the HTST system consists of a suite of NLP tools and includes a number of lexical knowledge data and disambiguation training model files. Hence, it is a resource-intensive software. In particular, when the HTST starts up, it costs some time to load all the resources into the system memory. Currently on average it takes 7308 milliseconds (ms) to start up if the VARD tool is not activated, and 8,917 ms if VARD is included, creating an issue of speed when the system processes corpus data stored in many files.

In order to alleviate the speed issue, a service wrapper was developed for the HTST, so that the HTST only needs to load the resources once, then interacts with the input data through a computer socket. When a pipeline mode is needed, client software of the HTST service can be included as a component in the pipeline, which drastically improves the speed compared to directly including the original HTST package. Such a design also improves the modularity of the system, as the HTST package can be maintained separately, and it allows loose coupling of software components, achieving a high level of separation of concerns.

To examine the impact of such service-based design on the HTST performance, we carried out an evaluation of speed on three sample corpus data of different sizes selected from the Hansard Corpus (cf. Section 6.1) on a PC (Intel® Core^{TM} i7 CPU Q 740 @ 1.73 GHz ×8) installed with Ubuntu 14.04 LTS 64-bit OS. Table 9 shows the results of the test. The second column lists the number of files containing the test data. Because the test data is contemporary English text, VARD is not activated in this experiment.

As shown in Table 9, when the HTST was used as a standalone tool that starts up for every individual test data file, it only achieved an average speed of 11.34 tokens per second. The speed for the biggest test data was increased because the data contained some larger files, which decreased the number of HTST start-ups. On the other hand, when the HTST was run as a service (a client program is used to call the service on the same PC), the processing times obtained improvements of 11.75, 11.86 and 5.87 times over the standalone package for the three test data, with an average of 9.83 times. With regard to the number of tokens processed per second, the service-based HTST processes 84.74 more tokens per second. Considering that the default HTST package runs with an average speed of 11.34 tokens per second, this is a substantial improvement.

It should be noted that, if the tools process a single large file, the speed improvement would be negligible. But in many practical applications, annotation tools need to process data in many batches of files, and in such cases the service oriented design can bring a significant impact on the speed of resource-intensive tools like the HTST. In the SAMUELS project, Hadoop was used to distribute the tagging process across a cluster of fifteen commodity machines. This configuration was able to tag the 1.26bn words of the Hansard Commons corpus in approximately six days, with a rate of 2436 tokens per second.

Table 9
Impact of service-oriented design on the speed of the HTST.

Tokens in test data	Number of files	HTST as standalone tool		HTST as service		Process time improvement	Increased Tokens/s
		Time (ms)	Tokens/s	Time (ms)	Tokens/s		
1000	17	137,129	7.29	11,669	85.70	11.75 times	78.45
10,000	165	1316861	7.59	111,041	90.06	11.86 times	82.47
100,000	593	5222849	19.15	889,205	112.46	5.87 times	93.31
Average	N/A	N/A	11.34	N/A	96.07	9.83 times	84.74

7. Conclusion

In this paper, we presented the HTST system, which is both the first semantic tagger that employs the *Historical Thesaurus of English*, a large historical thesaurus produced by linguistic experts, and the first to be trained on word sense definitions from the Oxford English Dictionary. This enables an unprecedented level of semantic annotation, and facilitates a deeper semantic analysis of language data compared to existing annotation tools. In particular, with the HT's rich information about word usage and the historical semantic development of the English language, the HTST is well adapted for annotating historical English corpus data since it is able to annotate words with historically sensitive categories to reflect their meaning at the time of use. With the ability to adapt its annotation given the date of the text, the HTST advances the capability of existing corpus annotation tools.

While the HTST has demonstrated an encouraging performance, there is still room for further improvement. We will integrate and develop better algorithms to fully exploit the information provided by the HT knowledge base to improve word sense disambiguation. We will also build more training data for enhancing the context-based disambiguation methods.

As it stands, the HTST has already been used for processing corpus data on a large scale, such as the Hansard Corpus, which provides valuable resources for research communities (Alexander et al., 2015a, 2015b; Alexander and Davies 2015; Demmen et al., preparation). We envisage that, with further development and improvement, the HTST will provide a powerful tool for both corpus-based studies and ICT applications.[4]

Acknowledgment

This work was supported by the UK Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (ESRC) as part of the Semantic Annotation and Mark-Up for Enhancing Lexical Searches (SAMUELS) Project (grant number AH/L010062/1).

References

Alexander, M., Baron, A., Dallachy, F., Piao, S., Rayson, P., 2015a. Metaphor, popular science and semantic tagging: Distant reading with the historical thesaurus of English. Digital Scholarship Humanit. 30 (1), 16–27.

Alexander, M., Baron, A., Dallachy, F., Piao, S., Rayson, P., Wattam, S., 2015b. Semantic tagging and early modern collocates. In: Proceedings of The Corpus Linguistics 2015 Conference. Lancaster University, UK, pp. 8–10.

Alexander, M., Davies, M., 2015. The Hansard Corpus 1803-2005. http://www.hansard-corpus.org (accessed 6.07.16).

Allan, J. (Ed.), 2012. Topic Detection and Tracking: Event-Based Information Organization. vol. 12, Springer Science & Business Media.

Anderson, W., Hough, C., Kay, C., Bramwell, E., Aitken, B., Hamilton, R., Alexander, M., 2015. Metaphor map of English. http://www.glasgow.ac.uk/metaphor (accessed 6.07.16).

Archer, D., McEnery, T., Rayson, P., Hardie, A., 2003. Developing an automated semantic analysis system for Early Modern English. In: Archer, D, Rayson, P., Wilson, A., McEnery, T. (Eds.), Proceedings of the Corpus Linguistics 2003 Conference, Lancaster University, UKpp. 22–31.

Archer, D., Rayson, P., Piao, S., McEnery, T., 2004. Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. In: Williams, G., Vessier, S. (Eds.), Proceedings of the Eleventh EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, FranceVolume III, pp. 817–827.

Balossi, G., 2014. A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves. John Benjamins, Amsterdam.

Baron, A., Rayson, P., 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In: Proceedings of the Postgraduate Conference in Corpus Linguistics. Birmingham, UK. Aston University. 22 May 2008.

Baron, A., Rayson, P., 2009. Automatic standardisation of texts containing spelling variation: How much training data do you need? In: Proceedings of the Corpus Linguistics 2009 Conference. Lancaster University, UK.

Baron, A., Rayson, P., Archer, D., 2009. Word frequency and key word statistics in historical corpus linguistics. Anglistik: Int. J. Eng. Stud. 20 (1), 41–67.

Chitchyan, R., Sampaio, A., Rashid, A., Rayson, P., 2006. Evaluating EA-Miner: Are early aspect mining techniques effective? In: Proceedings of Towards Evaluation of Aspect Mining (TEAM 2006). Workshop co-located with ECOOP 2006 (European Conference on Object-Oriented Programming), Nantes, Francetwentieth ed. pp. 5–8.

[4] For a demo website of the HTST tagger, see http://phlox.lancs.ac.uk/ucrel/semtagger/english; A client GUI tool of the HTST tagger is available at http://www.glasgow.ac.uk/samuels/

Crystal, D., Crystal, B., 2002. Shakespeare's Words; A Glossary and Language Companion. Penguin, London http://www.shakespeareswords.com (accessed 6.07.16).

Cunningham, H., Maynard, D., Bontcheva, K., 2011. Text Processing With GATE. Gateway Press, CA.

Demmen, J., Jeffries, L., Walker, B. (In Press). Charting the semantics of labour relations in House of Commons debates spanning two hundred years: A study of parliamentary language using corpus linguistic methods and automated semantic tagging. In: Kranert, M., Horan, G. (Eds.) '*Doing Politics': Discursivity, Performativity and Mediation in Political Discourse*. John Benjamins, Amsterdam.

Doherty, N., Lockett, N., Rayson, P., Riley, S., 2006. Electronic-CRM: A simple sales tool or facilitator of relationship marketing? The Twenty-Nineth Institute for Small Business & Entrepreneurship Conference. International Entrepreneurship—from local to global enterprise creation and development, Cardiff-Caerdydd, UK.

EEBO (Early English Books Online), 2003-2017. ProQuest LLC. http://eebo.chadwyck.com/home (accessed 22.05.17).

Fauconnier, G., Turner, M, 2002. The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York.

Gacitua, R., Sawyer, P., Rayson, P., 2008. A flexible framework to experiment with ontology learning techniques. Knowl. Based Syst. 21 (3), 192–199.

Garside, R., Smith, N., 1997. A hybrid grammatical tagger: CLAWS4. In: Garside, R., Leech, G., McEnery, A. (Eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman, London, pp. 102–121.

Greenblatt, S., Cohen, W., Howard, J.E., Maus, K.E. (Eds.), 1997. The Norton Shakespeare.

Hancock, J.T., Woodworth, M.T., Porter, S., 2013. Hungry like the wolf: A word-pattern analysis of the language of psychopaths. Legal Criminol. Psych. 18 (1), 102–114. doi: 10.1111/j.2044-8333.2011.02025.x.

Hendrickx, I., Marquilhas, R., 2011. From old texts to modern spellings: An experiment in automatic normalisation. JLCL 26 (2), 65–76.

Iacobacci, I, Pilehvar, M.T, Navigli, R, 2015. SENSEMBED: learning sense embeddings for word and relational similarity. In: Proceedings of the Fifty-Third Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing, Beijing, China, July 26−31, 2015, pp. 95–105.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I. (Eds.), 2009. Historical Thesaurus of the Oxford English Dictionary. Oxford University Press, Oxford. http://www.glasgow.ac.uk/thesaurus.

Kay, C., Roberts, J., Samuels, M., Wotherspoon, I., Alexander, M. (Eds.), 2016. Historical Thesaurus of EnglishUniversity of Glasgow, Glasgow. http://www.glasgow.ac.uk/thesaurus/. First published in print as *Historical Thesaurus of the Oxford English Dictionary*, 2009. Oxford University Press, Oxford. See also http://www.oed.com/ (last accessed 6 July).

Klebanov, B.B., Diermeier, D., Beigman, E., 2008. Political Analysis 16 (4), 447–463. doi: 10.1093/pan/mpn007.

Lehto, A., Baron, A., Ratia, M., Rayson, P., 2010. Improving the precision of corpus methods: The standardized version of early modern English medical texts. In: Taavitsainen, I., Pahta, P. (Eds.), Early Modern English Medical Texts: Corpus Description and Studies. John Benjamins, Amsterdam, pp. 279–290.

Levandowsky, M., Winter, D., 1971. Distance between sets. Nature 234 (5), 34–35. doi: 10.1038/234034a0.

Markowitz, D.M., Hancock, J.T., 2014. Linguistic traces of a scientific fraud: the case of Diederik Stapel. PLoS ONE 9 (8), e105937. doi: 10.1371/journal.pone.0105937.

McArthur, T., 1981. Longman Lexicon of Contemporary English. Longman, London.

Miwa, M., Thompson, P., Ananiadou, S., 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics 28 (13), 1759–1765.

Nakano, M., Hasegawa, Y., Nakadai, K., Nakamura, T., Takeuchi, J., Torii, T., Tsujino, H., Kanda, N., Okuno, H.G., 2005. A two-layer model for behavior and dialogue planning in conversational service robots. In: Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), Edmonton, Alta, Canada. doi: 10.1109/IROS.2005.1545198.

Ooi, V., Tan, P., Chiang, A., 2007. Analyzing personal weblogs in Singapore English: The Wmatrix approach. Studies in Variation, Contacts and Change in English. Volume 2. Research Unit For Variation, Contacts and Change in English (VARIENG).

Piao, S., Rayson, P., Archer, D., McEnery, T., 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. Comput. Speech Lang. 19 (4), 378–397. doi: 10.1016/j.csl.2004.11.002.

Potts, A., Baker, P., 2013. Does semantic tagging identify cultural change in British and American English? Int. J. Corpus Linguist. 17 (3), 295–324 Project Gutenberg. Online ebook resource. https://www.gutenberg.org/ (accessed 19.04.16).

Rayson, P., Archer, D., Piao, S., McEnery, T., 2004. The UCREL semantic analysis system. In: Proceedings of the Workshop on Beyond Named Entity Recognition Semantic Labelling for NLP Tasks in Association with Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugalpp. 7–12.

Rayson, P., Archer, D., Baron, A., Culpeper, J., Smith, N., 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on early modern English corpora. In: Davies, M., Rayson, P., Hunston, S., Danielsson, P. (Eds.), Proceedings of The Corpus Linguistics 2007 Conference. UK. University of Birmingham. 27−30 July 2007.

Rizzo, G., Troncy, R., 2012. NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In: Proceedings of the Demonstrations at the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics, pp. 73–76.

Roberts, J., Kay, C., Grundy, L., 1995. A Thesaurus of Old English. (King's College London Medieval Studies XI.). Second ed. Rodopi, Amsterdam 2000.

Semino, E., Demjen, Z., Demmen, J., Koller, V., Payne, S., Hardie, A., Rayson, P., 2015. The Online Use of Violence and Journey Metaphors by Patients with Cancer, as Compared with Health Professionals: A Mixed Methods Study. online edition BMJ Supportive and Palliative Care. doi: 10.1136/bmjspcare-2014-000785.

Taiani, F., Grace, P., Coulson, G., Blair, G., 2008. Past and future of reflective middleware: Towards a corpus-based impact analysis. The Seventh Workshop on Adaptive and Reflective Middleware (ARM'08) 1 December 2008, Leuven, Belgium. collocated with Middleware 2008.

Volk, M., Ripplinger, B., Vintar, S., Buitelaar, P., Raileanu, D., Sacaleanu, B., 2002. Semantic annotation for concept-based cross-language medical information retrieval. Int. J. Med. Inf. 67 (1-3), 97–112.

Vossen, P., 1998. EuroWordNet: Building a multilingual database with wordnets for European languages. The ELRA Newsletter 3 (1), 7–10. http://vossen.info/docs/1998/elra.pdf (accessed 22.05.17).

Weston, J., Bordes, A., Yakhnenko, O., Usunier, N., 2013. Connecting language and knowledge bases with embedding models for relation extraction. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 133–1371.

Scott Piao is a senior research associate of the School of Computing and Communications at Lancaster University, UK. He has rich experience in corpus tool development and natural language processing. He has worked on eight projects funded by EPSRC, ESRC, AHRC, EU and JISC, covering research topics of corpus construction and annotation, text mining, and application of corpus and natural language processing techniques in social computing. In the Semantic Annotation and Mark-Up for Enhancing Lexical Searches Project (SAMUELS, Ref. AH/L010062/1), he was the leading developer of the HTST semantic annotation software system, which is reported in this paper.

Fraser Dallachy is a research associate in English Language and Linguistics at the University of Glasgow. He is currently working on the AHRC-funded project, The Linguistic DNA of Modern Thought (AHRC grant AH/M00614X/1), a collaboration between the Universities of Sheffield, Glasgow, and Sussex.

Alistair Baron is a lecturer in the School of Computing and Communications at Lancaster University, UK. His primary research areas are Natural Language Processing and Cyber Security, with a particular focus on developing solutions that utilize language analysis techniques for combatting cybercrime. He previously researched and developed the VARD software which assists in normalizing the spelling variation found in historical, online, and learner texts.

Jane Demmen is a senior research associate in the Linguistics and English Language Department at Lancaster University, UK. She has worked on several corpus linguistics projects involving semantic annotation and tagging, including the project ‘Is There a Baron in the Commons’ (University of Huddersfield), which was part of the AHRC- and ESRC-funded Semantic Annotation and Mark-Up for Enhancing Lexical Searches (SAMUELS) project (grant reference AH/L010062/1) and the ESRC-funded Metaphor in End of Life Care project (Lancaster University; grant reference ES/J007927/1). Her research interests include corpus linguistics, Early Modern English drama, political discourse, metaphor and health communication.

Steve Wattam is now working in the commercial world but up until recently he was a lecturer and research associate at Lancaster University, where he obtained his PhD on the subject of representativeness in corpus linguistics. His research interests include the application of methods from statistics, machine learning, and natural language processing to problems involving the human understanding of large bodies of text. He has worked on numerous projects on the subjects of corpus construction, sampling of online and social media data, and large-scale text annotation. Steve can be contacted through his website: http://stephenwattam.com.

Philip Durkin is deputy chief editor of the Oxford English Dictionary, and has led the dictionary's team of specialists in Etymology for the past sixteen years. He is the author of The Oxford Guide to Etymology (2009; paperback 2011) and of Borrowed Words: A History of Loanwords in English (OUP 2014; paperback 2015), and he is editor of The Oxford Handbook of Lexicography (2015).

James McCracken leads the technology team for the Oxford English Dictionary, working in data modeling, natural language processing, and machine learning.

Paul Rayson is a reader in Computer Science at Lancaster University, UK and Director of the UCREL interdisciplinary Research Centre which carries out research in corpus linguistics and natural language processing (NLP). A long-term focus of his work is the application of semantic-based NLP in extreme circumstances where language is noisy e.g. in historical, learner, speech, email, txt and other CMC varieties. His applied research is in the areas of online child protection, learner dictionaries, and text mining of historical corpora and annual financial reports.

Marc Alexander is Professor of English and Linguistics at the University of Glasgow and Director of the *Historical Thesaurus of English*. He works primarily on the study of meaning in English, with a focus on lexicology, semantics, and stylistics through cognitive and corpus linguistics. His publications are on a range of topics in the linguistics of English, generally using the *Historical Thesaurus*, and he has written on metaphorical construction, historical semantics, parliamentary discourse, corpus research infrastructures, color terms across the history of English, and psycholinguistic manipulation in detective fiction.