WORD FREQUENCY AND KEYWORD EXTRACTION

AHRC ICT Methods Network Expert Seminar on Linguistics
8 September 2006, Lancaster University, UK

Issues for Historical and Regional Corpora: First Catch Your Word

Christian Kay, *University of Glasgow, Scotland.*

Keywords
word frequency, semantic, ambiguity, variable spelling, corpora, homonymy, polysemy, text analysis, non-standard language.

Abstract
The Historical Thesaurus of English (HTE) is a semantic index to the *Oxford English Dictionary* (OED) supplemented by Old English materials published separately in *A Thesaurus of Old English* (TOE). Word senses are organised in a hierarchy of categories and sub-categories, with up to fourteen levels of delicacy. The material is held in a database and first steps towards Internet publication are being taken by an AHRC ICT Strategy Project creating searches for use in a range of humanities disciplines. The main problem which besets searching historical texts is that of variable spelling – the further one goes back in time, the worse it gets. Similar problems affect texts in non-standard varieties, as experience of the Scottish Corpus of Texts and Speech (SCOTS) demonstrates. Dictionary headwords lemmatize common variants but by no means comprehensively; an alternative may be a rule-based system which predicts possible spellings. Corpora have further problems in ambiguity caused by homonymy and polysemy. The paper will suggest ways of addressing these problems.

Introduction
My interest in what words can tell us about a text stems mainly from two electronic projects, the Historical Thesaurus of English (HTE)[1] and the Scottish Corpus of Texts and Speech (SCOTS).[2] I have a further involvement with the Linguistic and Cultural Heritage Electronic Network project (LICHEN), headed by Lisa-Lena Opas-Hänninen at the University of Oulu, Finland, which aims to collect and display languages of the circumarctic region.[3]

TOE and SCOTS
HTE is not yet complete, but already has a daughter project, *A Thesaurus of Old English* (TOE),[4] a conceptually organised thesaurus of the surviving vocabulary of Old English (OE, c.700-1100 A. D.). Both TOE and SCOTS, which are freely available over the Internet, make some attempt to deal with word frequency. Much of the vocabulary of OE is assumed not to have survived, and that which does is unlikely to be representative of the whole. Because of this, the editors made use of four flags, somewhat similar to the register labels in modern dictionaries. These are **o** indicating infrequent use, **p** for poetic register, **q** for doubtful forms, and **g** for words occurring only in glossed texts or glossaries. Unlabelled words are common by default. The database as a whole contains some 50,700 meanings, deriving from almost 34,000 different forms. Of these meanings, around 30% have one or more of the above flags attached to them, which is an indicator of the peculiar nature of the surviving OE vocabulary.[5]

The flags are held in a separate field in the TOE database and can be searched either individually or in combination, yielding information about both individual words and their distribution

over semantic categories. From the TOE search menu, the user can select a search on the flags, and then **p** for words occurring only in poetry. The results can be browsed, or a particular semantic field, such as Section 13 Warfare, can be chosen. The Warfare screen shows that a total of 457 out of 1450 headwords in this section (around 32%) are marked with a **p** flag, which is the highest proportion in TOE. A sizeable proportion of these (302 or about 20%) are also marked as rare words by the **o** flag.[6] The specialized nature of the vocabulary of this area of Old English is thus confirmed.

SCOTS demonstrates frequency in a different way. If we institute a search on the form *war*, we find that it occurs in 90 documents (17.08% of the published corpus), and that these in turn contain 303,093 words (38.83% of the corpus), while the form *war* itself occurs 344 times.[7] However, a glance at the citations will show that this is not the end of the matter, since at least three meanings of *war* are represented in the selection:

> They *war* first biggit ('they were first built', where *war* is the third person plural past tense form of the verb 'to be').
> His faither is away tae the *war* ('his father has gone to war', where it is a noun).
> Be *war* of inserting sic lang words hinmest in the line ('be wary of inserting such long words at the end of the line'. Here *war* is an aphetic form of the adjective 'aware' or 'wary'.).

Since these three happen to be different parts of speech, grammatical parsing, which we have not yet tackled, would contribute to a solution here, but in general terms, despite many proposed solutions, multiple meaning remains a challenge for corpus searching. The joker in the pack is the third example, which is from a modern lecture text but is quoting from an *Essay on Poesie* written by James VI and I and published in 1585. Such temporal displacement is not unusual in texts. Direct quotation can be dealt with by tagging, and possibly ignored when results are returned, but there remains the problem of allusion, where a word may trigger reference to an event, a text, or whatever, necessitating access to a knowledge base. Perhaps this takes us further than tagging should have to go.

Where a word has only one meaning, figures such as those above can give us an idea of its relative frequency in the corpus. In a case like *war*, they can still provide useful pointers – for example, what form of the verb is used by those who do not select *war*? *Were* (most likely in written Scottish Standard English) or *was* (predictable for Glasgow and elsewhere in the central belt)? For any word, the extensive SCOTS metadata can answer questions which are crucial for interpreting frequency statistics as opposed to merely recording them. Where do the users of a form come from? What is their sociolinguistic profile? Do they use the word primarily in speech or in writing? In which genre does the word occur?

From my point of view, however, that of someone interested in historical and regional language, the title of this seminar, 'Word Frequency and Keyword Extraction', is really the wrong way round. Issues of frequency can only be tackled if we are confident that we are able to retrieve the material we need from our corpora. At least two problems currently stand in the way of such confidence: the problem of variable spelling and, as just demonstrated, the problem of semantic ambiguity. Such problems affect not only linguists but also the wide range of humanities scholars engaged with regional or historical texts.

Spelling

To take spelling first, any work on historical texts has to solve the problem of the variations in spelling which occurred in English until at least the eighteenth century, and which become more marked the further back in time one goes. Even in modern standard varieties there is a degree of variation, as in the resistance of British English writers to using '-ize' forms in words like 'realise', and one might also want to capture erroneous spellings. A parallel problem involves spelling variation between varieties and sub-varieties of English. In some of these, such as Scots, there may be no generally-accepted written standard, with orthographic choice being left to individuals or groups, who may not themselves be consistent.[8] This problem is shared by the many languages in the world which have

underdeveloped written forms. In the LICHEN project, we will be tackling languages with virtually no written form, such as the Finnish varieties Meänkieli and Kven, but that presents different problems.

Several methods have been used to deal with spelling variation. The simplest is to present the user with an alphabetized concordance of all the words in a corpus, allowing them to search on likely variants.[9] Such a method presupposes a well-informed user, otherwise variants which are not alphabetically close or obvious, such as *fit* as a variant of *what* in north-eastern Scots, may be missed. A more foolproof method is to lemmatize the variants by tagging while the corpus is under construction, thus building up a spelling dictionary for that particular body of texts.[10] This method is likely to be successful, but at considerable cost in human effort, and depends on skilled annotators being available. Moreover, its success cannot be guaranteed beyond the selected texts. A more sophisticated procedure involves extending the use of wildcard searches by writing algorithms predicting the range of possible spellings, either overall or for particular periods or varieties. This is not an easy task, either linguistically or computationally, but if the method were sufficiently generalizable, it would prove invaluable in many humanities disciplines, since it would allow scholars to import and search texts of their own choosing rather than rely on prepared corpora. For the next generation of tools, we should perhaps be looking to such frameworks: in many areas of the humanities, electronic texts are relatively easy to acquire, but annotated corpora are not. A case in point was provided at the Digital Resources in the Humanities Conference 2005 (DRH), where two historians interested in trade and material culture in the early modern period discussed the issues involved in setting up an electronic *Dictionary of Traded Goods and Commodities 1550-1820* based on a digitized corpus of primary materials.[11] Problems were encountered in retrieving the very variously spelled terms for items such as foodstuffs, spices, and dyes in their database.

A good deal of information about spelling in the history of English is lemmatized under the headwords in the *Oxford English Dictionary* (OED), where the main variants are given century by century.[12] These listings suggest a way forward, but also illustrate the extent of the problem, as a glance at the OED data for two homophonous English words, *peace* and *piece*, demonstrates.

> *Peace*, noun
> Forms: 2-4 pais, 2-6 pes, (3-5 pays, peys, 3-6 peis, 4 payes, 4-5 payse, pese, pees, Sc. and north. pess), 4-6 pece, (5 peese), 5-6 peas, pease, (pesse, Sc. peice, 5-7 peax, 6 Sc. peiss, pace), 6- peace.
> *Piece*, noun
> Forms: 3-7 pece (3-5 pees, 4 pise, 4-5 pice, peis, 5 pes, peyce, peese, 5-6 pes(s, pesse); 5- piece, (5 pyece, 5-8 peace, 6 pease, peise, peyss, (Sc. peax), pysse, 6-7 peece, 6-8 peice).

(The figures show the centuries of currency of particular forms; thus ‘2-4’ indicates twelfth to fourteenth century. Sc = Scots)

It would be possible to use these listings as a starting point, instructing a program to search for all variants. However, this would only be partially helpful since the spellings under the headwords are the most common variants rather than a comprehensive listing, and other forms might well occur. An alternative, as discussed above, would be to attempt to predict the range of possible variants. Some of the resulting algorithms would be very broad, as for the range of vowel variation above, while others would be quite restricted. In the sixteenth to eighteenth centuries, for example, *ph* could be substituted for *f* in words of classical origin such as *phantastic*, *phrentic* (frantic), or *phanatic*. As an added complication, the ending of such words could be *-ic, -ick, -ik, -icke, -ike*, or even *-ique.* These endings, however, have widespread substitutability over a longer period in a greater range of words, and so could be encapsulated in a more general rule. Indeed, it may be that we could make progress by ignoring the vowels and concentrating on the consonants: the *Dictionary of Old English* (DOE) manages to find variously-spelled phrases by treating all vowels as one in their searches.[13]

Ambiguity

Even if an adequate system for retrieving spellings is devised, we will still face the problem of semantic ambiguity. Although neatly distinguished in modern usage, the spellings of many homophones overlap historically, as in the *peace/piece* example above, and the researcher could not always be sure which one was being retrieved. For a common word such as *piece*, with many meanings and spellings, manual sifting (e.g. of the 7340 hits returned for *piece* in OED2) would be extremely time-consuming.

This situation raises the more general problem of disambiguating multiple meaning in what are traditionally distinguished as homonyms (*peace/piece* coming from different roots) and polysemes (the seventeen main meanings of *piece* listed by the OED, not to mention sub-senses, phrases and compounds). This distinction, and especially the role of polysemy in the extension of meaning through such processes as metonymy and metaphor, is of considerable interest in semantics even if computationally irrelevant. Disambiguation of such forms has long been a challenge in Natural Language Processing (NLP), with increasing success being achieved through projects such as WordNet,[14] MindNet,[15] and the preference-based approach of Wilks' Pathfinder project at New Mexico State University.[16] Such work exploits the compositionality of lexical meaning and the tendency of words to co-occur with others from the same semantic domain. In the case of Wilks and his colleagues, there is a refreshing and demonstrated appreciation of the contribution of information in published dictionaries to creating NLP tools.[17]

Historical Thesaurus of English (HTE)

This project is effectively a semantic index to the OED, supplemented by Old English materials from TOE. HTE's projected 650,000 word meanings are presented in 26 major categories, each arranged in a detailed semantic taxonomy of up to twelve hierarchical places, thus showing the position of each meaning within the overall structure. Every section has an explanatory heading in modern English, which can be traced back through the hierarchy to create a definition – indeed, it would be possible to reverse the process and to produce a unique kind of structured dictionary with the headwords in alphabetical order.

Each section is organized internally in chronological order, with words retrievable through a unique number in the 29-field database. Someone searching for *piece* in HTE would find that it occurs at various times in numerous categories, such as land, bread, drugs, people, armaments, games, etc., while *peace* occurs in flowers, absence of war, freedom from care, absence of noise, etc. First steps towards Internet publication of the materials are being taken by an AHRC ICT Strategy Project creating datasets and searches, including delimitation of words by dates of currency, for use in a range of humanities disciplines.[18] In addition to being of interest to linguists, the organization of vocabulary in semantic categories can cast light on such topics as the development of material culture, social organization, and intellectual pre-occupations.[19]

Thesauri have long been employed as a component of automatic parsing tools, as in the use of Tom McArthur's *Lexicon of Contemporary English*[20] in Lancaster's USAS package (UCREL Semantic Analysis System), which is being redeveloped to cope with historical factors in the sixteenth to eighteenth centuries.[21] A favourite for such research has been Roget's well-known *Thesaurus of English Words and Phrases*,[22] which was used in early work of this kind.[23] An interesting current example of a historically-focused use of Roget occurs in research by Terry Butler of the University of Alberta to map a tagged version of the notebooks of the English poet Samuel Taylor Coleridge (1772-1834) to the categories of the first edition of Roget (1852), thus creating a contemporary subject index.[24]

Although the overall structure of HTE was originally devised by a componential analysis of key OED definitions,[25] it is essentially a folk taxonomy, more akin to McArthur than Roget. In the example below, 03 represents the major class 3.Society (the other two being 1.The Physical Universe, and 2.The Mind), with 03.03 giving the most general words for the concept of armed hostility. Old English words are given first, marked simply 'OE', but earliest and latest dates of currency from the OED are given from 1150 on. An entry like 'win<(ge)winn OE - c1275' gives the post-OE form, followed by its OE ancestor, with a last recorded use of around 1275. 'Conflict 1611—' indicates a first recorded date of 1611, with continuous currency into present day English. Interrupted currency or scarcity of

examples is indicated by the use of a plus sign between dates. OED labels such as ‘poetic’ or ‘dialectal’ can be downloaded if desired.

> 03.03. n Armed hostility: geflit OE, garniþ OE, guþ OE, hild OE, niþ OE, orlege OE, orlegniþ OE, sæcc OE, unfriþ OE, unsibb OE, win<(ge)winn OE - c1275, camp<camp OE - c1400, cock a1300, battle a1300—, arms c1374—, armour 1387 - 1602, pugny 1456, hostility 1531—, combattencie 1586, conflict 1611—, hostilities 1613—

This paragraph is followed by a sequence of semantic sub-categories, then by parallel categories for other parts of speech. Sub-categories read back to the main heading: in those below, ‘armed hostility’ must be supplied after the preposition, giving ‘outbreak of armed hostility’, etc., as the full heading.

> 03.03. /01. n (.outbreak of):
> 03.03. /02. n (.declaration of):
> 03.03. /03. n (.commencement of):

Nineteen major categories follow 03.03, starting with 03.03.01 War, and moving through Battle, Victory, Defeat, Warriors, Weapons, and so on until we finally reach 03.03.19 Peace/absence of war. Degrees of subordination within sub-categories are represented by the number of dots. The example below shows a pathway, reading from the lowest level, defining a person or ship carrying / a flag of truce / as part of a suspension of hostilities / leading to their cessation / and thus peace.

> 03.03.19. n Peace/absence of war:
> 03.03.19. /06. n (.cessation of hostilities):
> 03.03.19. /06.01. n (..suspension of hostilities):
> 03.03.19. /06.01.03. n (...flag of):
> 03.03.19. /06.01.03.01. n (....person/ship carrying):

Both the dates and the semantic structure displayed here could be used in creating a probability-based method of disambiguating historical word forms. One could predict that if the word *peace* or a variant occurred in a context where other words from that HTE category also occur, then it is likely to be OED sense I.1.a. ‘Freedom from, or cessation of, war or hostilities; that condition of a nation or community in which it is not at war with another’, or one of its sub-sections, that is involved, rather than the peace rose or a piece of cake. If the approximate date of the target text is known, then only words contemporary with that text need be considered. The matching would not be 100% successful, since unknown words and unexpected contexts are bound to occur. Rules would have to be developed for the amount of context and levels of hierarchy required, bearing in mind that the level of semantic delicacy of HTE is much greater than that of most thesauri. Music, for example, has 7,471 meanings under 2,416 category headings, while Animals has 29,883 meanings and 12,818 headings.

Formal novelty is, of course, linked to spelling variation. Anyone in the seventeenth century could produce the forms *fantastic*, *fantastick*, *fantastik*, *fantasticke*, *fantastike*, or *fantastique*, not to mention *fantastickal*, *fantastikal*, *fantastical*, or *fantastiqual*; if actual occurrences of any of these have escaped the OED net, they could nevertheless be recognized as potential words. Common types of metonymic extension could be incorporated into a search tool, such as the name of a tree being used for its wood or its fruit (e.g. *apple*). Thesauri also reveal metaphoric extension; if there is a noticeable overlap in words between an abstract and a concrete category (as in Anger/Heat), then there is often a metaphorical connexion, with the potential for new metaphors to be added to the set. Overall, HTE could be a useful addition to electronic resources for historical text linguistics, including data-mining of older texts – but only if the spelling problem is solved first.

Keywords

An HTE category or subcategory can also contribute to work on keywords, in the sense of words that reveal cultural preoccupations. Although frequency is not marked as such, absence of labels such as 'rare' or 'dialectal' indicates general currency. A long date range indicates likely importance over a period of time, while the semantic clustering of many words may indicate a significant concept. New ideas or technology may be represented by a sudden spurt of words at a particular period. Anyone who wonders about the relative importance of *war* and *peace* as talking points in English might reflect on the relative numbers of words for these concepts in HTE, as shown below. Within the *War* category there are strikingly long lists for military artefacts and personnel – the material goods may change, but the concept unfortunately lingers on. Within *Peace*, there is very little.

	Records	Headings	Words
War	16785	3885	12900
Peace	406	101	305

Conclusion

Electronic dictionaries and corpora are now familiar resources in humanities computing, useful both in linguistic research and in the many disciplines where searching for words can produce historical or social information or literary insights. We are reaching a point where these tools are moving on, becoming more complex in their computing architecture and more powerful in what they can achieve. It is to be hoped that before too long the considerable achievements of NLP in tools for text analysis can be harnessed for the benefit of work on historical and non-standard language.

Notes

[1] A research project in progress at the Department of English Language, University of Glasgow, see <http://www.arts.gla.ac.uk/sesll/englang/thesaur/homepage.htm>

[2] <http://www.scottishcorpus.ac.uk>: John Corbett, Jean Anderson, Christian Kay, and Jane Stuart-Smith, funded by AHRB grant B/RE/AN9984/APN17387.

[3] See Juuso, Ilkka, Anderson, Jean, Anderson, Wendy, et al, 'The LICHEN Project: Creating an Electronic Framework for the Collection, Management, Online Display, and Exploitation of Corpora', in Hardie, A. (ed.), *Digital Resources for the Humanities Conference 2005 Abstracts*, 27–29.

[4] Roberts, Jane, and Kay, Christian, with Grundy, Lynne, *A Thesaurus of Old English* (London: King's College London Medieval Studies XI, 1995), (2nd edition, Amsterdam: Rodopi, 2000). An electronic version, supported by British Academy grant LRG-37362, can be seen at: <http://libra.englang.arts.gla.ac.uk/oethesaurus>

[5] For more detail, see Christian Kay, 'A Thesaurus of Old English Online', <http://www.oenewsletter.org/OEN/reports.php>

[6] These figures will fluctuate slightly, since TOE is periodically updated on receipt of new materials from the Toronto *Dictionary of Old English* project, <http://www.doe.utoronto.ca>

[7] These figures will vary, since SCOTS is updated at regular intervals. It has a target of four million words, 20% spoken, by the end of current funding in spring 2007.

[8] A list of recommended spellings, based mainly on frequency of occurrence, is being drawn up by Scottish Language Dictionaries Ltd (SLD), the body with responsibility for developing academic dictionaries of Scots. SLD has, of course, no power to impose its use.

[9] An example of this kind of approach is the corpus of *Middle English Medical Texts* (MEMT) compiled by Irma Taavitsainen and her team at the University of Helsinki, and obviously aimed at sophisticated users (published by Benjamins of Amsterdam on CD). (personal communication; see also <http://www.eng.helsinki.fi/varieng>)

[10] Anneli Meurman-Solin of the University of Helsinki has used this method in preparing her forthcoming *Tagged Corpus of Scottish Correspondence*. (personal communication; see also <http://www.eng.helsinki.fi/varieng>)

[11] See Cox, Nancy, and Dannehl, Karin, 'The Rewards of Digitisation: A Corpus-Based Approach to Writing History', in Hardie, A. (ed.), *Digital Resources in the Humanities Conference 2005 Abstracts*, 13–14..

[12] Simpson, John A. (ed.) *The Oxford English Dictionary (OED) Online* (Oxford: Oxford University Press, March 2000 –), <http://www.oed.com>.

[13] <http://www.doe.utoronto.ca>

[14] <http://wordnet.princeton.edu> (accessed 8/8/05).

[15] <http://research.microsoft.com/nlp/Projects/MindNet.aspx> (accessed 8/8/05).

[16] For a discussion of developments from the 1950s onwards, see Wilks, Yorick A., Slator, Brian M., and Guthrie, Louise M., *Electric Words: Dictionaries, Computers and Meanings* (Cambridge, MA, and London: MIT Press, 1996). More recent work, including the MALT project (Mappings, Agglomerations and Lexical Tuning), is described at <http://www.dcs.shef.ac.uk/~yorick>(accessed 1/10/05).
[17] For a defence of dictionaries in linguistic terms, see Kay, Christian, 'Historical Semantics and Historical Lexicography: will the twain ever meet?', in Coleman, Julie, and Kay, Christian (eds), *Lexicology, Semantics and Lexicography in English Historical Linguistics: Selected Papers from the Fourth G.L. Brook Symposium* (Amsterdam: Benjamins, 2000), 53–68.
[18] Smith, Jeremy, Horobin, Simon, and Kay, Christian, Lexical Searches for the Arts and Humanities, AR112456.
[19] See, for example, Kay, Christian, 'Historical Semantics and Material Culture', in Pearce, Susan M. (ed.), *Experiencing Material Culture in the Western World* (London and Washington: Leicester University Press, 1997), 49-64.
[20] McArthur , Tom, *Lexicon of Contemporary English* (London: Longman, 1981).
[21]Archer, D., McEnery, T., Rayson, P., and Hardie, A., 'Developing an Automated Semantic Analysis System for Early Modern English', in Archer, D., Rayson, P., Wilson, A., and McEnery, T. (eds), *Proceedings of the Corpus Linguistics 2003 Conference* (Lancaster: UCREL, 2003), 22–31. See also <http://www.comp.lancs.ac.uk/computing/research/ucrel/usas> (accessed 2/10/05)
[22] Roget, Peter Mark, *Thesaurus of English Words and Phrases* (London: Longman, 1852 and subsequent editions).
[23] Wilks et al, 1996, 130–131, citing Masterman, M., 'The Thesaurus in Syntax and Semantics', *Mechanical Translation*, 4 (1957), 1–2.
[24] *Improving Access to Encoded Primary Texts*, ACH-AHRC abstract <http://mustard.tapor.uvic.ca/cocoon/ach_abstracts/xq/xhtml.xq?id=170> (accessed 19/7/05).
[25] Kay, C., and Samuels, M. L., 'Componential Analysis in Semantics: its Validity and Applications', *Transactions of the Philological Society* (1975), 49-81.