Proceedings Paper:
Elliott, JR and Atwell, ES (2001) Visualisation of long distance grammatical collocation patterns in language. In: Banissi, E, (ed.) Proceedings of the 5th International Conference on Information Visualisation. 5th International Conference on Information Visualisation, 25 - 27 Jul 2001, London, UK. IEEE Computer Society , 297 - 302. ISBN 0769511953

Visualisation of Long Distance Grammatical Collocation Patterns in Language

Contact author: John Elliott, jre@comp.leeds.ac.uk

Co-authors: Eric Atwell, eric@comp.leeds.ac.uk
Bill Whyte, billw@comp.leeds.ac.uk

Organisation: Centre for Computer Analysis of Language and Speech,
School of Computing, University of Leeds, Leeds, Yorkshire, LS2 9JT England

Abstract

Research in generic unsupervised learning of language structure applied to the Search for Extra-Terrestrial Intelligence (SETI) and decipherment of unknown languages has sought to build up a generic picture of lexical and structural patterns characteristic of natural language. As part of this toolkit a generic system is required to facilitate the analysis of behavioural trends amongst selected pairs of terminals and non-terminals alike, regardless of which target natural language was selected. Such a tool may be useful in other areas, such a lexico-grammatical analysis or tagging of corpora. Data-oriented approaches to corpus annotation use statistical n-grams and/or constraint-based models; n-grams or constraints with wider windows can improve error-rates, by examining the topology of the annotation-combination space. We present a visualisation tool to help linguists find "useful" PoS-tag combinations, and cohesion between linguistic annotations at other levels; and suggest some possible applications.

1. Introduction: language identification in unknown signals

Research on NLP applied to the Search for Extra-Terrestrial Intelligence (SETI) has sought to build up a generic picture of lexical and structural patterns characteristic of natural language. (Elliott et al [5,6,7]) describe algorithms and software developed to characterise and detect generic intelligent language-like features in an input signal, using Natural Language Learning techniques: looking for characteristic statistical "language-signatures" in test corpora. As a first step towards such species-independent language-detection, a suite of programs has been developed to analyse digital representations of a range of data, and use the results to extrapolate whether or not there are language-like structures, which distinguish this data from other sources, such as music, images, and white noise. It is assumed that generic species-independent communication can be detected by concentrating on localised patterns and rhythms, identifying segments at the level of characters, words and phrases, without necessarily having to "understand" the content. Furthermore, the simplifying assumption is made that a language-like signal will be encoded symbolically, i.e. some kind of character-stream. A language-detection algorithm for symbolic input can use a number of statistical clues: data compression ratio, "chunking" to find character bit-length and boundaries, and matching against a Zipfian type-token distribution for "letters" and "words". SETI researchers do not claim extensive (let alone exhaustive) empirical evidence that such language-detection clues are "correct"; the only real test will come when the Search for Extra-Terrestrial Intelligence finds true alien signals. If and when true SETI signals are found, the first step to interpretation is to identify the language-like features, using techniques like the above.

2. Correlation profiles

An intermediate research goal is to apply Natural Language Learning techniques to the identification of "higher-level" lexical and grammatical patterns and structure in a linguistic signal. We have begun the development of tools that measures the correlation profile between pairs of words, parts of speech, and potentially other linguistic labels in a tagged corpus, as a precursor to deducing general principles for 'typing' and clustering into syntactico-semantic lexical classes. Linguists have long known that collocation and combinational patterns are characteristic features of natural languages, which set them apart [13]. Speech and language technology researchers have used word-bigram and n-gram models in speech recognition, and variants of PoS-bigram models for Part-of-Speech tagging. In general, these models focus on immediate neighbouring words, but pairs of words may have bonds

despite separation by intervening words; this is more relevant in semantic analysis, eg [14, 4]. We sought to investigate possible bonding between type tokens (i.e., pairs of words or between parts of speech tags) at a range of separations, by mapping the *correlation profile* between a pair of words or tags. This can be computed for given word-pair *type* (w1,w2) by recording each word-pair *token* (w1,w2,d) in a corpus, where d is the distance or number of intervening words. The distribution of these word-pair tokens can be visualized by plotting d (distance between w1 and w2) against frequency (how many (w1,w2,d) tokens found at this distance). Distance can be negative, meaning w2 occurred *before* w1.

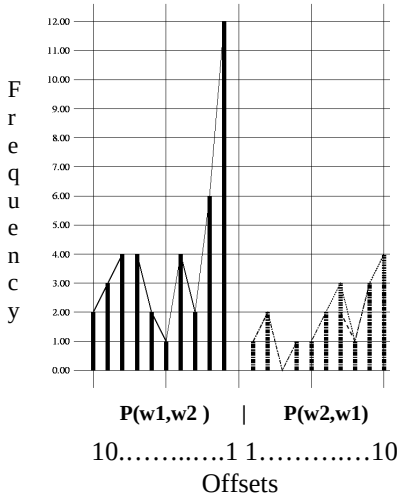

Figure 1. Visualisation of a correlation profile for a word pair (w1=the,w2=king)

Figure 1 shows the results for the relationship between a content and function word, so identified by looking at their cross-corpus statistics. More detailed examples and explanation are below. It can be seen that the function word has a high probability of preceding the content word but has no instance of directly following it. At least metaphorically, the graph can be considered to show the 'binding force' between the two words varying with their separation. We are looking at how this metaphor might be used in order to describe language as a molecular structure, whose 'inter-molecular forces' can be related to part-of-speech interaction and the development of potential semantic categories for the unknown language.

So far we have mainly been working with English, but we have begun to look at languages which represent their functional relationships by internal changes to words or by the addition of prefixes or suffixes. Although the process for separating into functional and content terms is more complex, we believe the fundamental results should be consistent.

3. Applications in Part-of-Speech Tagging

Part-of-Speech tagging programs have tried to combine statistical n-gram-based models with local combinational constraint rules in various ways. The first large-scale PoS-tagging system TAGGIT [8], used to PoS-tag the Brown Corpus, used a set of tag-combinational constraints "hand-crafted" by linguists, where tag combination preferences were specified within a window or local context of 5 words. In practice, researchers found most of the constraints, which fired, relied on only the immediately preceding or following word. So, the successor PoS-tagger built for tagging the follow-on LOB Corpus, CLAWS [11] instead used a 1st-order Markov or bigram model of tag-coocurrence, learnable from a pre-tagged training corpus (sampled from Brown), augmented with some hand-picked longer-context constraints when post editors thought these could improve accuracy [1,2]. The widely-used Brill tagger [3] uses constraint rules rather than a statistical bigram model, but these constraints are machine-learnt from a pre-tagged training corpus, so the system can learn different tagging schemes from different tagged training corpora [15]. The ENGCG approach [10] requires an expert linguist to devise a Constraint Grammar rule-set using linguistic knowledge and corpus evidence. Others have extended the statistical bigram approach to more sophisticated statistical models (eg Manning and Schutze [12]); however these may require larger training sets, and furthermore the increase in sophistication, eg from bigrams to trigrams, does not yield corresponding increase in accuracy: error rates have not improved dramatically.

It appears that many trigrams are "redundant" in that they would not alter the tagging decision from that made by the simpler bigram model. A general observation is that a model based on bigrams or immediate neighbours, whether Markovian or constraint-based, can go a long way, but lower error-rates can be achieved most efficiently by selective use of longer-context patterns or constraints only when necessary. A tool to visualize the topology of the combinational-space for Part-of-Speech tags may help linguists in their search for useful or significant combinations: the mapping may show up peaks and troughs, which correspond to longer-distance combinational constraints.

4. The Toolkit

The process by which samples are analysed is sub-divided into a number of separate processes, each within a separate program, which are “piped” at the command line. This method is used to enable command line arguments and facilitate choice of output formats.

At the command line two main sets of arguments can be supplied to constrain the analysis. The first of these is simply which pair of tokens are required for analysis. The other constrains the range [window size] for analysis and later display. Potentially, the system can depict the behaviour of selected units over the entire sample in a single snapshot. However, for our purposes and usually most practical applications, analysis needs to be restricted to a closer neighbourhood, where language is most likely to possess conditional grammatical relationships.

Once the data has been 'cleaned' it is passed onto programs, which ‘flag’ and record the positional information of selected command-line arguments within the input data for subsequent behavioural analysis. This positional information is segmented according to the selected window size.

Using this constraint, it scans for the first argument; once found it then looks for an instance of the second argument and records the distance [offset] it was located at. If no instance is found of the second argument within the range selected, a value of zero is recorded and the next instance of the first argument is sought.

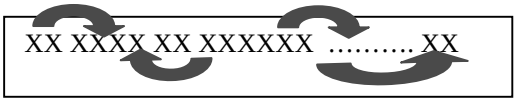

Figure 2. Bilateral co-occurrence detection

As the reverse behaviour of the selected arguments has the potential to be useful in developing a complete picture, the program also records the behaviour where argument one follows argument two. This analysis of 'mirrored pairs is performed concurrently providing time efficiency at run-time.

$$f(w1) \,.\, f(w2) = \sum_{x>0} f(w1, x, w2) + \sum_{x<0} f(w2, x, w1)$$

and the probability w1 precedes w2 at offset is:

$$f(w1) \qquad f(w1, x, w2)$$
$$f(\forall w) \qquad f(w1, n, w2)$$

Left hand side probability = $f(x) / \sum_{x<0} f(x)$

Left hand side probability = $f(x) / \sum_{x>0} f(x)$

The constrained accumulative behaviour for the cohesive bonding of the chosen linguistic objects are then collated and formatted for visualisation.

5. The output.

Finally, the data extracted is passed to the last link in the chain, which prepares the information for display. Here, two options are available. One is in the form of a graph, which shows the two sets of information - arg1..arg2 and arg2..arg1 - either side of a centre line . This style of representation is perhaps preferable when first searching for 'trends' prior to any precise analysis supplied numerically.

The alternative display is numeric, where the individual frequencies, independent and conditional probabilities are displayed for interpretation. Here again, both sets of ordered argument pairs are displayed, with the additional indicator to whether the frequency of the pairs' occurrence at that particular word separation is greater or less than that of the combined independent probabilities.

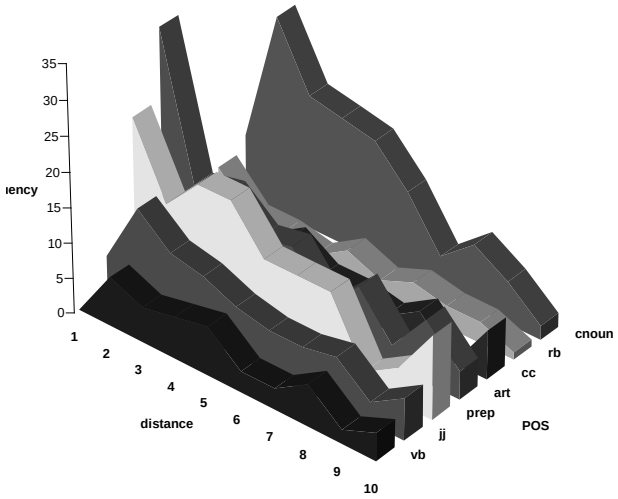

Figure 3: VB-tag profile

In addition to measuring the frequency of exclusive collocations within a given window size, the system also measures and includes the frequencies of all second arguments for each offset within range. These are made distinguishable by representing each combination using different colours and texture (for black and white displays). Therefore, for this measurement, the system does not cease to look for subsequent occurrences of argument 2 once the first is found, moving on then to the next occurrence of argument1 but continues to record all second argument occurrences up until the

given limit. This particular measure has proved useful in observing grouping of parts-of-speech.

To enable analysis of multiple selections and how they compare with each other, the information extrapolated is then ported for 3D graphical representation [see fig 3]. This particular stage will eventually be integrated for purposes of efficiency but is not essential.

Examining language in such a manner also lends itself to summarising the behaviour to its more notable features when forming profiles. Therefore conducting information compression akin to Principal Component Analysis. A technique more usually found in applications conducting analysis of images and found to be extremely effective.

6. Finding long-distance combinational constraints

Using a five thousand-word extract from the LOB corpus [9], a number of parts-of-speech pairings were analysed for their cohesive profiles. The arbitrary figure of five thousand was chosen, as it both represents a sample large enough to reflect trends seen in samples much larger (without losing any valuable data) and a sample size, which we see as at least plausible when analysing ancient or extra-terrestrial languages where data is at a premium. Figure 4 shows a sample of the main syntactic behavioural features for their co-occurrence ranging over the chosen window of ten words.

Fig 4	noun	adjective	adverb	prep
noun	β, λ_3	δ^*	λ_2	β^*
adj	β^*	β	$\delta, \lambda_{5,9}$	λ_2
adverb	Z, λ_5	λ_7	β	β^*
Prep	δ^*, λ_2	λ_2	δ^*, λ_7	δ, λ_3
conj	$\delta^*, \lambda_{3:4}$	β	β, Z_6	λ_4
Verb	λ_2	λ_2	β	β^*
article	β^*	β^*	$\delta, \lambda_{3,8}$	Z, λ_2

Fig 4	conj	verb	article
noun	β^*, λ_6	δ, λ_2	δ^*, λ_2
adj	$\lambda_{2,4}$	δ	$\delta^* \lambda_3$
adverb	δ, λ_9	β	λ_2
Prep	λ_3	Z^*, λ_9	β
Conj	Z	λ_5	β^*
Verb	δ, Z_9	Z	β^*
Article	Z^*	Z	Z, λ_4

Figure 4. Analysis of distinguishing grammatical collocations of main Parts of Speech.

Most of the combination patterns found correspond to tag-bigrams, which could be extracted automatically in a Markov model. However, some longer-distance cohesive trends were found, indicated by λ_n in the above table, where n is the offset distance. For example, in our sample, adverbs (Rb) were never immediately followed by a common noun (Cnoun), but there was a peak at a separation of 5.

Such information could be used to guide development of Constraint Grammars. The English Constraint Grammar described in [10] includes constraint rules up to 4 words either side of the current word (see Table 16, p352); the peaks and troughs in the visualisation tool might be used to find candidate patterns for such long-distance constraints.

7. Assessing syntactic separation of related word-classes

In addition to the above general overview of word-tag combinational topology, the visualisation tool can focus on specific subsets of the grammar. WH-words are given a detailed sub-classification in the LOB tag-set, with five different W-tags (tags for WH-words, starting W…). The W-tag set from the LOB corpus was analysed to ascertain whether such fine-grained lexical sub-categories are justifiable. In the table below [figure 5], principle behavioural elements of the syntactic topography show results for each major W-tag followed by one of a number of selected major PoS-tags (the reverse being also automatically calculated) using a larger sample set of over eighty five thousand words.

Fig 5	Idem	noun	adj	adverb
Wpr	Z	$\delta, \lambda_{3:5}$	δ, λ_3	λ_2
Wrb	Z	$\delta, \lambda_{2,3}$	$\delta, \phi_{2:10}$	β, λ_{10}
Wdtr	Z	$\delta, \lambda_{3,4}$	δ	λ_2
Wdt	Z	δ, λ_2	δ	Z
Wp	Z	Z, λ_2	Z	Z

Fig 5	prep	conj	verb	article
Wpr	$\delta, \lambda_{3,4}$	$Z, \lambda_{4,6,8}$	β	δ, λ_4
Wrb	$\delta, \lambda_{2:6}$	$\delta, \lambda_{4,5,7}$	$\lambda_{2,3}$	β
Wdtr	$\delta, \lambda_{2:5}$	$Z, \lambda_{6:10}$	λ_2	$\beta, \lambda_{2:5}$
Wdt	$Z, \lambda_{3,4}$	Z	ϕ	$\beta, \lambda_{4,5}$
Wp	Z	Z	Z	Z

Figure 5. Analysis of grammatical collocation patterns of LOB WH-tags illustrated in fig 6.

Key:
Z = Zero bigram - or at offset specified - occurrences.
δ = Very weak bonding - near zero - at bigram occurrences.
β = Strong bonding at bigram co-occurrences.
* = Indicates opposing cohesive trend when P.O.S. reversed.
λ_n = High peak beyond bigram at offset distance of 'n'
ϕ = Flat distribution across offsets – bigram bonding evident.

Figure 6 shows the topography of one example W-tag set obtained from corpora analysis. The tables indicate broadly similar combinational behaviour, with very little distinctive behaviour to justify such sub-categorisation. There may well be a case for combining most, if not all of these tags into one grammatical category; the subclasses are lexical classes rather than syntactic classes.

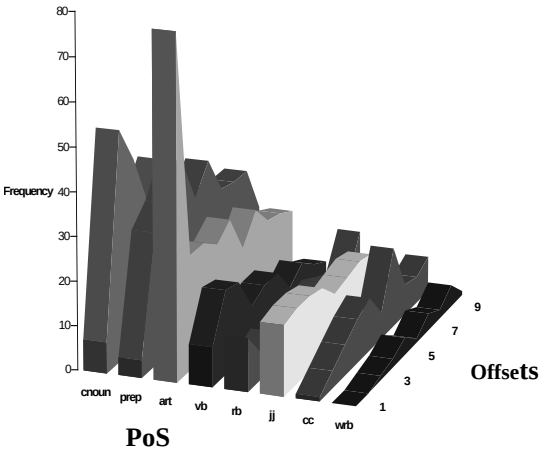

Figure 6: WRB-tag profile

To investigate whether particular combinations display distinguishable traits at more distant separations, which may further aid unsupervised language learning, a one hundred word/parts-of-speech tag window was employed.

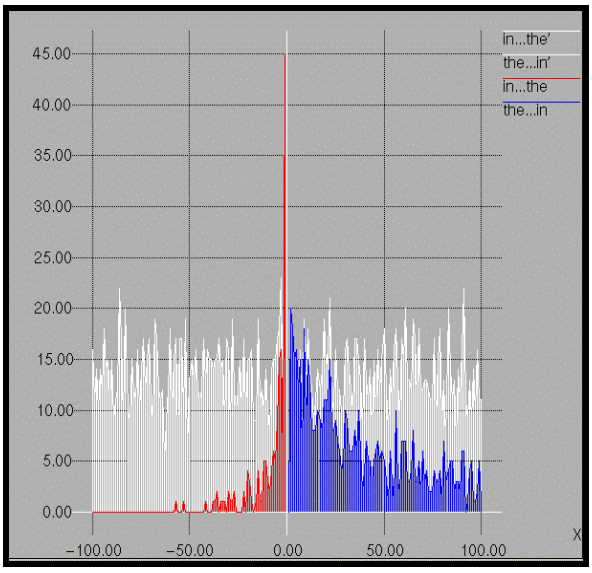

Figure 7: in/the profile

The rationale here being: given no prior knowledge except those gleaned from previous stages of unsupervised analysis, can statistically based features of the annotation-combination space topology contribute towards clustering the functional words into parts of speech.

To ascertain the feasibility of such a hypothesis, English functional words, which are discovered during previous unsupervised analysis, were analysed. It was found, using such distant behaviour, that frequent - almost bound - word combinations such as in figure 7 below, display a marked 'tailing off' in the direction where bonding is evident, in contrast to the opposing direction where repulsion occurs at the immediate zero offset bigram occurrence.

The red area depicts behaviour for the bigram in/the, whilst the blue depicts its opposite. This profile can be compared with the non-bound word pair topology of 'in/of', seen in figure 8.

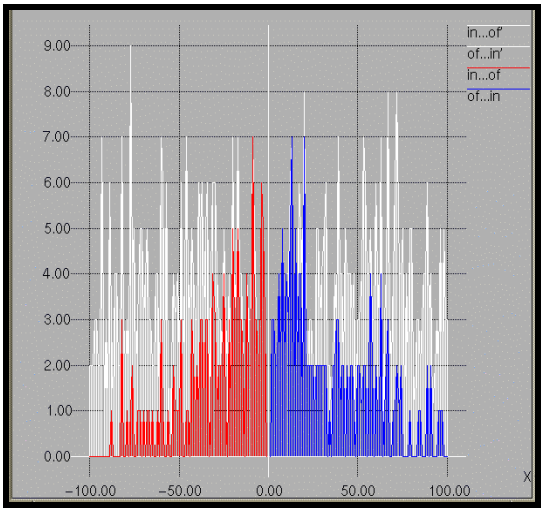

Figure 8: in/of profile

The white areas, which contribute towards the overall topology, are where the exclusivity of selected pair combinations are not enforced and intervening secondary occurrences of a selected word are ignored.

8. Devising constraint grammars at other linguistic levels

Corpus linguists are developing corpora annotated at higher linguistic levels, for example discourse speech-act labeling. A "grammar" of speech acts could be developed in the style of the Helsinki Constraint Grammars, by building up a rule-set of combinational constraints. The visualisation tool could be applied to any level of corpus annotation, in building a constraint-based description of tag-combination at that level.

9. Conclusion

In general, we realise that testing our language detection algorithms will be a significant issue. We do not have examples that we know to be definitely from non-human, but intelligent origins, and we need to look extensively at signals of non-intelligent origin which may mimic some of the language characteristics described above. This will form a significant part of our future work and we welcome discussion and suggestions.

This is not to claim that the system described is revolutionary or indeed unique. However, we are not aware of any system that tackles the problem of terminal and non-terminal behaviour in such a visual and flexible way. Manning and Schütze state [12], " *Any technique that lets one visualise the data better is likely to bring to the fore new generalisations and to stop one from making wrong assumptions about the data*". We therefore submit that the ability to visualise interactive behaviour over more distant offsets in a single snapshot has the potential to identify often overlooked but nonetheless important behaviour.

Potential implications for tagging and probability assignments underlying Hidden Markov Models are the trends observed at offsets, which exceed often more commonly used conditional information such as trigrams. Using this long-distance view, which is enhanced with visual representation, we believe a more intuitive and immediate feel for conditional behaviour can be gleaned and a more complete and reliable model developed.

Once we have a clearer picture of how visualisation peaks and troughs correspond to constraint rules, it should also be possible to semi-automate the process of extracting constraint grammars from (tagged) corpora. Plans are in place to completely automate this system for public domain on the Internet.

10. References

[1] Atwell, Eric. 1982. *LOB Corpus tagging project: post-edit handbook.* Department of Linguistics and Modern English Language, University of Lancaster.

[2] Atwell, Eric. 1983. *Constituent Likelihood grammar*. ICAME Journal 7:34-66.

[3] Brill, Eric. 1993. *A Corpus-based approach to language learning.* PhD thesis, Department of Computer and Information Science, University of Pennsylvania.

[4] Demetriou, George, and Atwell, Eric. 2001. *A domain independent semantic tagger for the study of meaning associations in English text.* To appear in proceedings of IWCS4: Fourth International Workshop on Computational Semantics, Tiltburg, Netherlands.

[5] Elliott, John, and Eric Atwell. 2000. *Is there anybody out there? : The detection of intelligent and generic language-like features.* In Journal of the British Interplanetary Society, 53:1/2, 13-22.

[6] Elliott, John, Eric Atwell and Bill Whyte. 2000. *Language identification in unknown signals.* In Proceeding of COLING'2000, 18th International Conference on Computational Linguistics, pages 1021-1026, Association for Computational Linguistics (ACL) and Morgan Kaufmann Publishers, San Francisco.

[7] Elliott, John, and Eric Atwell 1999. *Language in signals: the detection of generic species-independent intelligent language features in symbolic and oral communications* in Proceedings of the 50th International Astronautical Congress, paper IAA-99-IAA.9.1.08, Amsterdam, Netherlands.

[8] Greene, Barbara and Gerald Rubin. 1981. *Automatic grammatical tagging of English*. Providence, R.I.: Department of Linguistics, Brown University.

[9] Johansson, Stig, Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB corpus: users' manual*. Bergen University, Norway: ICAME, The Norwegian Computing Centre for the Humanities.

[10] Karlsson, Fred, Atro Voutilainen, Juha Heikkila, and Arto Anttila. 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text.* Berlin: Mouton de Gruyter.

[11] Leech, Geoffrey, Roger Garside and Eric Atwell. 1983. *The automatic grammatical tagging of the LOB corpus*. ICAME Journal 7:13-33.

[12] Manning, Christopher and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

[13] Sinclair, John. 1991. *Corpus, concordance, collocation, describing English language*. Oxford University Press.

[14] Wilson, Andrew and Paul Rayson. 1993. *The automatic content analysis of spoken discourse*. In C Souter and E Atwell (eds), Corpus based computational linguistics. Rodopi, Amsterdam.

[15] Atwell, Eric, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter, and Sean Wilcock. 2000. *A comparative evaluation of modern English corpus grammatical annotation schemes.* ICAME Journal, 24: 7-23. Bergen: International Computer Archive of Modern and medieval English.