Old spellings, new methods: automated procedures for indeterminate linguistic data

Hugh Craig and R. Whipp
Centre for Literary and Linguistic Computing, School of Humanities and Social Science, The University of Newcastle, Australia

Correspondence:
Hugh Craig,
Centre for Literary and Linguistic Computing,
School of Humanities and Social Science,
The University of Newcastle,
Callaghan, NSW 2308,
Australia.
E-mail:
hugh.craig
@newcastle.edu.au

Abstract

The authors have worked over several years on a software tool to make word counts from an archive of old-spelling early modern English plays and poems. In this article we present the outcome, a computational model for dealing automatically with variant spelling, implemented in an application which we call an 'Intelligent Archive'. We also reflect on the perspective on Early Modern English, and on the probabilistic aspect of language in general, gained from working through the practical problems which arose in establishing the model.

1 Counting words in old spelling

The authors have worked over several years on a system to make reliable frequency tables for all the different words in an archive of old-spelling early modern English plays and poems, focusing on the period 1580–1640. In this article we present the outcome, a computational model for dealing with variant spelling, implemented in a software application, which we call an 'Intelligent Archive'. We also reflect on the perspective on Early Modern English, and on the probabilistic aspect of language in general, gained from working through the practical problems which arose in establishing the model.

Our underlying research interest is in computational stylistics, and in authorship attribution in particular. A key method here is to count frequencies of the words that a given author tends to use more than the norm, as a guide to the attribution of a contested text. This means collecting instances of a given dictionary word across all the texts regardless of its local spelling. The corpus we use currently contains around 180 plays and 50 poems and poem collections, yielding 3.8 million running words. The texts are all based on early printed versions in original spelling. We also need accurate counts of words in these old-spelling texts for calculating measures like type token ratios, for examining collocations, and for concordances. In fact most applications are best served by the underlying dictionary word rather than the surface form.

Nothing seems more straightforward than recognizing a word for counting purposes. A word, as every reader knows, is a string of characters separated by a space or by punctuation. Indeed, the fact that words are easy to identify is a prime reason to choose them as units of analysis. Confronted by a sixteenth-century or early seventeenth-century printed volume, however, complexities arise. Punctuation, capitalization and font changes are not going to affect word counts, so can be ignored, but beyond that immediate decisions are needed. Spelling to the modern eye is erratic, and there are numerous irregular contractions and unfamiliar printer's marks. A typographically exact transcription of the early printed original would include

Literary and Linguistic Computing, Vol. 25, No. 1, 2010.
doi:10.1093/llc/fqp033 Advance Access published on 12 October 2009

long 's' forms as well as the now standard short 's', marks like tildes to indicate a missing 'm' or 'n', and double 'v' letters to represent 'w'. At another level, on the boundary between typographical and spelling variation, there are the familiar substitutions of 'u' for 'v', 'i' for 'j', 'i' for 'y', and vice versa. There are also unpredictable substitutions of 'ie' for 'y' at the ends of words and apparently random additions or omissions of a trailing 'e'.[1]

Once decisions have been made about representing typographic forms, there are contractions and separated forms to deal with, and possessives, which are irregularly represented in many of the texts, so that a trailing 's' may indicate a simple plural, a singular possessive, a plural possessive or indeed, in some contexts, an elided form of 'is' or 'has'.[2]

Once we get beyond typography to the level of spellings as more conventionally understood, we find some that are difficult to resolve into modern forms even for the experienced editor working case by case, let alone by an automated process (Bowers, 1966, 1975, Wells, 1979, Salmon, 1986). Some words which are now orthographically distinct were, as best we can tell, used interchangeably in early modern English, with various spellings. Examples are *travel* and *travail*, *human* and *humane*, *diverse* and *divers*, *wrack* and *wreck* and *mettle* and *metal* (Wells, 1979, pp. 11–12, Bowers, 1966, pp. 159–61).[3] In early modern English these were single words with two senses but a single pronunciation and a shared set of spellings, rather than two words as in modern English (de Grazia, 1990, pp. 154–5). Wells says Shakespeare uses *mettle* and *metal* in a 'significantly ambiguous' way in *Julius Caesar* (Wells, 1979, p. 11). These present what Wells calls 'insoluble problem[s]' for modernization in a modern-spelling text (Wells, 1979, p. 12), and equally for a countable old-spelling version. The researcher preparing a countable text shares another problem with the editor of a modernized version, i.e. distinguishing variant spellings from separate though obsolete forms. The modernizing policy for the second Arden Shakespeare series, for instance, was 'to preserve all older forms that are more than variant spellings' (qtd. in Bowers, 1975, pp. 292–3 n. 8). Yet there is no sure guide to what is a 'doublet' or variant spelling and what is a distinguishable form. Is 'murther' a separate, archaic form of 'murder', for instance, or one of its spelling variants? What about 'banket' and 'banquet'? Wells notes that the obvious authority, the *OED*, is inconsistent in distinguishing 'spellings' from 'forms' (Wells, 1979, p. 7).

In the context of a mechanical process to make word-counts these protean forms of words merge into a larger class of ambiguous spellings, where the words involved are clearly separate, but overlap in orthography. 'Hart,' for instance, appears as a variant of *hart* (a male deer) and also as a variant of *heart*. This means that without some further disambiguation we must discard both *hart* and *heart* for counting purposes, since we cannot be sure of an accurate count of either. Similarly, 'weeke' is used in printed texts in the period for the modern words *weak*, *week* and *wick*.

The question then arises, why not work with versions already modernized by an editor and published? The first answer is that while most plays from the period have now been transcribed in old spelling in *Chadwyck-Healey Literature Online* and in *Early English Books Online*, modernized editions are much rarer and by no means freely available for research. Shakespeare may be available in electronic modernized forms, and the better known plays of writers like Jonson, Middleton and Marlowe, but we may wait a long time for public-domain modernized electronic versions of the plays of Robert Wilson, or William Haughton, let alone little-read anonymous plays. Old-spelling electronic editions are what is available now to the researcher.

Where a modern edition exists, and is in the public domain, it is important to know whether the standardization of forms follows the same protocols as other modernized plays in the archive. The ideal for counting purposes is a modernized edition of each play, completed on consistent protocols, and where each local modernization has been carefully considered by an expert editor. (Ideally, also, the orthography as printed would be preserved in a tagged format like XML, in case researchers wish to recover the original spelling, in order to study authorial spellings, for instance, or re-interpret it according to new understandings or new requirements.) A modernization project of this kind

would represent immense labour and at the very least a large delay in compiling any sizeable corpus. To have a large quantity of texts available now for statistical work based on counting words, one must use old-spelling texts, and resort to machine aid to deal with variant spelling. The question is one of scale: half a dozen plays, for instance, could be prepared by hand to a high degree of accuracy from old spelling texts in a reasonable amount of time. For work with a much larger corpus, one with some claims to representativeness of authors, genres, and periods, we need machine help.

Other reasons have been advanced for working in old spelling. Typically, as Bowers notes, modernized printed texts are based on a preceding edition, whereas 'the old-spelling edition must refer directly back to the originals' (Bowers, 1966, p. 155). Something analogous happens with electronic versions, where taking over a modernized version would mean adopting blindly all its editorial decisions and its errors.

In terms of language more generally, scholars championing old-spelling texts have sometimes suggested that the wildly varying spelling of the early modern period reflects a different linguistic sensibility from the modern one (Andrews, 1998, de Grazia and Stallybrass, 1993). In the early modern texts we find 'a semantic field that, not yet ruled by lexical statute, accommodates verbal vagrancy' (de Grazia and Stallybrass, 1993, p. 264). In terms of sheer variation, it is remarkable that Shakespeare's six surviving signatures spell his surname five different ways (Schoenbaum, 1981, pp. 94–7). De Grazia and Stallybrass list fourteen different sixteenth- and seventeenth-century spellings of *one* (de Grazia and Stallybrass, 1993, p. 262, n. 26). Another indication of the proliferation of spellings is in the fact that in one set of three folio pages of a play manuscript, scholars have counted sixteen spellings that are unmatched in the whole Chadwyck-Healey collection of English literature from 900 to 1900 (Jackson, 2007, paragraph 1.19). Thus it would seem that the variety of spelling we find in printed works in the Early Modern period, which is already spectacular, would pale beside the variety used in handwritten papers.

With the sort of fluid, 'unanchored' spelling we find in the early modern texts, words can merge into one another and the spelling forms may on occasion reflect deliberate punning and plays on words. De Grazia and Stallybrass quote the Folio version of Macbeth's speech on the vision of Banquo and his descendants (IV.1.113–5):

> . . . And thy haire
> Thou other Gold-bound-brow, is like the first:
> A third, is like the former. (de Grazia and Stallybrass, 2003, p. 265)

They suggest that there is in the Folio's 'haire' 'a phonetic and semantic convergence of what is for us three distinct words—*air*, *heir*, and *hair*'. They go on, 'However phantasmagoric this kind of semantic slipping and sliding may seem to a modern sensibility, Renaissance textuality encourages it' (de Grazia and Stallybrass, 2003, p. 265).

Variant spelling also allows for more specific punning allusions to related words. Ron Rosenbaum quotes a line from the 1604 *Hamlet* Quarto: 'The ayre bites shroudly, it is very colde' (Rosenbaum, 2006, pp. 251–2). The allusion to the shroud enfolded in the spelling is remarkably apt for the occasion, which is Hamlet's meeting with the Ghost. It is possible to be over-ingenious in detecting significant word-play through variant spelling, however. Plays are predominantly performances and it is hard to believe that the spellings 'haire' or 'shroudly' would be heard, unless, as G. B. Evans suggests, variant spellings suggest variant pronunciation (qtd in Wells, 1979, p. 4). Otherwise the potential play on words in (say) 'shroudly' could only be realized for an audience if *shroud* and *shrewd* were pronounced alike in Shakespeare's day. It is difficult to decide that question with any certainty (Gurr, 2001, Crystal, 2008, pp. 127–8). Richard Morton makes the general comment that 'There are few puns or witticisms dependent on spelling in the drama' (Morton, 1984, p. 113).

In any case, even if we accept recent arguments that some printed play texts represent versions prepared especially for readers (Erne, 2003), there is no guarantee that the spellings we have are authorial rather than those of the scribe or the compositor. Nevertheless, it is well to remember the possibility

that a particular spelling may be motivated by a wish to keep alive multiple semantic possibilities or to point the way to a particular etymology. Poems or prose works written to be read rather than performed and seen into print by the author may well use a spelling for 'complex effects' which are at least muffled by standardized orthography (Evans qtd. in Morton 113). Here any one-to-one software for spelling standardization is as guilty as the modernizing editor.

Faced with this perplexing linguistic indeterminacy, the researcher preparing a countable text cannot avoid difficult choices. For a countable old-spelling text we can think in terms of four levels of compression. The surface level is the orthographic word as it appears in the early versions; the standardized orthographic word represents a first consolidation of forms; then there is the level of the lemma (the dictionary headword); and beyond that aggregation into word classes or semantic groups. The second level is our target for the work presented here, and the focus for the transformations of the text into word counts which are described below. The first level reflects too many merely typographical vagaries, which would most likely have no relation to the spoken word on stage, and no necessary connection to an author. The third and fourth levels, on the other hand, lose too much potentially useful information. In support of the second level, the level of the standardized orthographic word, as opposed to the third, the level of the dictionary headword, is the finding of the Lexis Research project for the Office for Scientific and Technical Information (OSTI) that collocation patterns are specific to particular forms of words rather than to lemmas (Sinclair *et al.*, 2004, pp. xix).

Our approach with the lexical words is thus to work at the level of the standardized orthographic word: in terms of linguistic 'shallowness,' a level just below the text in its original form, since many typographical variations are ignored, and variant spellings are reduced to headwords wherever possible, but without any further analysis by part of speech, by sense, or by lemma. In our system *cry*, *cries*, *cried*, and *crying* are all headwords. Moreover, we do not differentiate between instances of *cry* as a noun and as a verb, essentially for pragmatic reasons: manual tagging would be too laborious, and automatic part-of-speech tagging working on old-spelling text would be too inaccurate.[4]

In editing terms the sort of countable text we prepare approximates a 'conservative reprint' of a chosen document, rather than a 'critical edition' (Bowers, 1975, pp. 293, 289). The latter would 'attempt to recover what the author actually wrote', by comparing the chosen copy-text with any other documents that had any authority, taking into account press variants like proof corrections, the characteristics of compositors, and so on (Bowers, 1975, pp. 290–1 n. 2, 292). This would require not only great labour but the exercise of much individual judgment, and would not fit the aim of producing counts that arise directly from a given early witness to the text. In Bowers' terms the countable text envisaged here is strictly an edition of a document, rather than of a text, representing the text 'only in its raw and inevitably corrupt state' (Bowers, 1975, pp. 291–3).

2 Our software for processing old-spelling texts

2.1 Function words

An underlying distinction between function words and lexical words informs the model: function words are much denser in distribution than lexical words on the whole, and thus differing statistical methods are suited to the two kinds of words. Using different tests with different variable sets makes for analyses which can corroborate each other, or not.

We begin therefore by standardizing the spelling and the separation of a set of function words manually, expanding contractions and ellipses and resolving non-standard separated forms. We work with a fixed list of 200 function words, comprised of a core group of very common function words and a supplementary group of words which serve to complete some component sets within the list, such as the pronouns, regardless of frequency. For these 200 words there is generally a small range of variations (*one*, which does appear in the list, is an exception).

These words themselves are a very limited set, given the tens of thousands of different words that appear in a large corpus. They are (predominantly) abundant, and evenly distributed; they constitute around 56% of all the word tokens used in the plays and poems. We use a Text Encoding Initiative format, so we can preserve the original version alongside the standardized one, as in '<reg orig="bee">be</reg>' or '<reg orig="it's">it is</reg>' or '<reg orig="al waies">always</reg>'. These operations are carried out with the help of 'find and replace' functions in word-processing software. Function words not included in the list of 200 are treated with the lexical words.

2.2 Lexical words: conflation by lists and internal spelling equivalents

Lexical words present different problems, as already discussed. Variant spellings of these words are too numerous, and too various, to 'find and replace' by hand. In terms of modern orthography, there are over 60,000 different word-forms in our current archive, and half of these occur only once. With 'bee' or 'it's', one can search for instances and replace them with the TEI tagging. Spelling variations with the lexical words, by contrast, are multiple and unpredictable. We need an automatic process, but one which mirrors the 'delicate decisions' Stanley Wells says are required for a hand-made modern-spelling edition (Wells, 1979, p. 4).

To deal with these problems we adopt a different approach. Instead of editing the texts, we work with an external library of word-forms, lists of spelling variants that can be associated with given modern spellings.[5] One source of these equivalents is the 'Spellings' listed in the *OED*. Oxford University Press allowed us to harvest these from a raw-text file of the entire *OED*. We also collected spelling variants by matching short passages from old-spelling Shakespeare plays with parallel passages in modernized versions in a public-domain Shakespeare edition (the *Moby Shakespeare*). The machine determined where a word in an old-spelling version was an equivalent for another word in a modernized one, based on the number of words common to the two passages in the same position, cross-checked against the target word's longest common subsequence (LCS). This provided another set of spelling variants. We also applied some replacement rules to account for typographical variations. We allowed 'u' always to be replaced by 'v', 'i' by 'j', and so on. We gathered just under 280,000 pairings between word-forms in all, spellings that are linked as orthographic variations on an underlying common lexical item, and then extended these through our replacement rules. This library is then the equivalent of the 'sense inventories' collected by Word Sense Disambiguation researchers from dictionaries, or from Wordnet (Agirre and Edmonds, 2006, p. 7).

We use the prose fiction collection in the British National Corpus as our reference for deciding which among these word-forms constitutes a modern spelling: if an orthographic form appears in this part of the BNC, it is for our purposes a modern spelling. There are some word-forms in our corpus that do not appear in the BNC collection, and that have not been linked through our lists with a modern-spelling form either. These 'orphan' words, together with the BNC modern-spelling forms, constitute our collection of 'headwords'; beside these we put the lists of spellings that are associated with a modern spelling, which make up our stock of 'variants'.

In this way we attempt to replicate through our software the steps that a researcher might take when searching old-spelling texts manually. MacDonald P Jackson describes a hand-crafted process like this for an authorship project using the Chadwyck-Healey Literature Online corpus (Jackson, 2002). Aiming to collect a complete set of search terms to cover all the spellings of given words, he drew on his experience from a lifetime of reading the texts, checked *OED* for any he might have missed, and included all the substitute letters habitually used by printers in the period. The website has a search engine with some wild cards and incorporating typographical variations, which allowed some short cuts. Any results he regarded as valid were compiled for a cumulative total.

Our counting of headwords proceeds in a parallel way, combining counts for any associated variant under the appropriate headword. Naturally, the reliability of the counts depends on the degree to

which our variant lists are accurate. Any spelling variant which we have missed, any variant, that is, that we have wrongly classified as a headword, and any spelling variant that is wrongly included in a list, introduces an element of error. This is a different, and larger, source of error than for the function words, where, given the manual editing and checking that takes place, counts will be wrong or dubious only through corruptions or inadvertent errors in a file, and through decisions in the case by case standardization which might be challenged.

As we have mentioned, resolving ambiguity in spelling variants is a close cousin to the well-established practice of Word Sense Disambiguation. It is interesting that WSD in its explicit form, in which instances are assigned to senses in a separate process, has not had many practical applications (Agirre and Edmonds, 2006, pp. 3, 11). In the present case the disambiguation is carried out as a discrete operation but does serve a practical purpose—collecting accurate counts of headwords in old-spelling texts—well.

In any analysis we do using the counts, then, we have to bear in mind that the base figures for lexical words and for function words will have two different kinds, and extents, of what Gabriel Egan calls 'indeterminacy' (Egan, 2005). The tests which rely on the lexical-words data will ideally employ as many word-variables as possible, so that errors tend to cancel each other out. We must always remember that not every variant of a headword will be included in a count, and the occasional variant of a different headword will find its way into the tally. One precaution which is more than ever important in this situation is to check the method with samples of known classification before trusting a result with mystery cases.

2.3 Resolving ambiguous variants by context

Some variants, however, appear in more than one headword list. For this group we need to use some further process to assign them to the correct candidate among the rival headwords. This task is closely related to word sense disambiguation (WSD), an important field within computational linguistics. The 'homograph' variety of WSD is concerned with 'identifying senses of homonyms that are not semantically related, such as *bank* (the land formation) and *bank* (the financial institution)' (Leacock and Chodorow, 1998, p. 266). This is in essence the same case as the ambiguous spellings 'weeke' and 'hart' already mentioned. The candidate senses—*week*, *weak*, or *wick*, and *heart* and *hart*—are at least as different in sense as are the two significations of *bank*. This is the simpler variety of WSD; the harder type involves disambiguating senses that are semantically related, such as '**serve** as a director' and '**serving** fruit as a salad' (Leacock and Chodorow, 1998, p. 268; bolding in original).

The method we developed depends firstly on contexts, like all WSD methods that use supervised or unsupervised learning, rather than information from other sources like dictionaries. It goes back to Warren Weaver's idea that, while it would be impossible to resolve ambiguity in sense for a single word in isolation, revealing a larger and larger window around it makes the identification of the correct sense progressively more certain (Weaver, 1955). Madhu and Lytle have already applied Bayes' theorem to this sort of problem (Madhu and Lytle, 1965).

The method we have devised draws on the prose fiction collection in the British National Corpus, already used as a source for modern spellings. This part of the BNC has around 16 million words. We gather information on the words which appear near each headword in the collection. Which words appear near *hart* and how often? Which words appear near *heart*? Then we can look at the words that in fact appear near our instance of 'hart' in its location in a play and see if the pattern is closer to a *hart* one or a *heart* one.

Thus for each instance in the old-spelling text we note the words that appear in the immediate context. We can then see how often these same words appear in the same positions in our extract from the BNC near the first headword that is linked to the target word-form. We then add up these scores to get a composite score for this first headword. We do the same for the second and any other headwords. We then choose the headword with the highest composite score, on the grounds that the words

Table 1 Resolving an ambiguous spelling in *Hamlet* I.v.55-8 (1604 version) by context

		'−3'	'−2'	'−1'		'+1'	'+2'	'+3'	
		Celestiall	**Bed**	**And**	**Pray**	**On**	**Garbage**	**But**	
		0/16m	7,502/16m	421,313/16m		115,552/16m	79/16m	96,617/16m	
Pray	369/16m	0	0	42		1	0	2	Score = 0.312
Prey	175/16m	0	0	1		9	0	1	Score = 0.478

neighbouring the target instance fit the pattern of this headword the best.

Research elsewhere in the comparable field of word sense disambiguation has shown that this kind of classifier performs as well as more complex techniques such as neural networks and content vectors (Leacock *et al.*, 1993). Local context as used in our classifier has emerged as more successful in WSD than a broader approach to detecting the topic of a given context (Yarowsky cited in Agirre and Edmonds, 2006, p. 3).

The disambiguation procedure we have designed is what Fernando Pereira calls a 'naïve method', in which 'the model probability estimates [are] just the relative frequencies of observed events' (Pereira, 2000, p. 1242). A more sophisticated method might begin with the difference in patterns in the contexts of rival headword candidates in the reference corpus, rather than with the particular context for the target ambiguous spelling. For each pair or group of rival headwords one could assemble a large collection of discriminating context words, each with a weighting, and a function based on these could form a classifier for disambiguation in particular cases. The computational resources needed for such an operation, given the thousands of ambiguous words encountered in larger corpora, would be considerable, however.[6]

2.4 An instance of disambiguation by context in practice

The Ghost in *Hamlet* reports in lurid detail on his murder by his brother, and excoriates his widow for betraying him and marrying his murderer. Finally he makes a disgusted general pronouncement about lust:

> So [lust], though to a radiant angel linked,
> Will [sate] itself in a celestial bed,
> And prey on garbage.
>
> But soft, methinks I scent the morning's air . . . (I.v.55–8)[7]

In the texts of Shakespeare's time, the spellings 'pray' and 'prey' were used both for *pray* as in 'pray for me' and *prey* as in 'birds of prey'. In fact, in the 1604 quarto of *Hamlet*, which is the version we use as our standard *Hamlet*, the spelling in this passage is 'pray'. For editors, as for those preparing a countable text, this presents a problem. Modern editions generally print 'prey' in their version of this passage, as with the Riverside version quoted above. Ron Rosenbaum argues that *pray* is a plausible reading and should be regarded as a substantive difference in the 1604 Quarto, not a spelling variant (Rosenbaum, 2006, pp. 96–8). Who is right? Is this an instance of prayer or of predation? How should it be spelled in a modernized edition, and where should it be assigned in a word count?

As already noted, a solution which is built in to the Intelligent Archive is to look at the immediate context for clues. The software checks with a large corpus (16 million words) in modernized spelling to see whether the context for the instance of 'pray' in the 1604 Hamlet is more like a typical *prey* context or a typical *pray* context.

It takes into account the three words before and the three words after the target word. Table 1 shows the results.[8]

Just concentrating on the word before, that is, *and*, we can see that in the corpus this word appeared 42 times just before *pray* and once just before *prey*. The Intelligent Archive takes into account the fact that *and* is a common word (over 421,000 occurrences in the corpus's 16 million), and therefore quite likely to appear by chance, and also that there are 369 instances of *pray* and 175 of *prey*, so it is really 42 appearances out of 369 and 1 out of 175. At this point 'pray' is ahead, i.e. the component score for this position is larger for 'pray' than for

'prey,' since 42 divided by 421,000 and again by 369 is larger than 1 divided by the same amounts. However, looking at the word following the target word, *on*, we see that this occurs in this position once out of 369 times in the case of 'pray' and nine times out of 175 times in the case of 'prey'. *On* itself is a less common word than *and* (there are 115,000 of them in the corpus as opposed to 421,000), so if we add the scores this more than counterbalances the evidence of *and* and puts 'prey' ahead overall. The other words do not contribute much—in fact only *but* has any occurrences in the corpus in the right slot—so 'prey' emerges the winner and the context part of the disambiguation module within the Intelligent Archive duly counts this instance in the 'prey' column.[9]

If we turn to the Corpus of Contemporary American English (COCA), 385 million words, we find similar relationships (Davies, 2008). There are 442 instances in all of 'pray' or 'prey' followed by 'on', and of these 47 are 'pray' and 395 'prey'. Other things being equal, therefore, a word which may be either 'pray' or 'prey' and is followed by 'on' is much more likely to be 'prey'. Curiously, three of the COCA instances of 'pray on' are clearly misspellings, indicating that the 'pray'/'prey' variant spelling has survived in modern American English, or has re-emerged there, and that the ratio should really be 44 to 398.

Another of the instances of 'pray on' in COCA illustrates the pressure to understand 'pray on' as 'prey on'. Here is the sentence quoted: 'The question of who gets to determine the destiny of the land, and of the people who live on it—those with the money or those who pray on the land—is a question that is alive throughout society.'[10] This is difficult to disambiguate with any certainty, though the opposition of 'those with the money' and 'those who pray on the land' suggests that the second group of citizens are the beneficiaries of rhetorical approval from the writer and are probably engaged in prayer rather than predation. Despite this (to the present reader at least) it takes an instance of 'prayers' in a following sentence of the original document to confirm decisively that the spelling is in fact correct (LaDuke, 1999, p. 5; in the corpus extract, the sentence is a quotation embedded in a different document).

A further extract from COCA, this time from a science fiction story from 2008, shows again that without further helpful context it is difficult to be sure that a [pre I] sound followed by an 'on' is an instance of *pray*:

> 'May the blood of your foes be thinned with tears,' the Sith replied cheerily. 'Perhaps we may pray together on Araso.' Sam wondered if he had meant 'prey' or 'pray.' Of course, with the Sith, they could be the same word.[11]

It may be, then, that a reference to probabilities by way of a corpus can help to make sense of a passage, by turning an intuitive language expectation into a quantifiable measure. A probabilistic assessment might well be a very useful additional source of guidance for editors of old-spelling texts.

One might also argue that both the *prey* and *pray* senses were intended by the writer in the *Hamlet* passage, and that the mind of the listener is meant to flicker between the two possibilities. The preceding words *celestial* and *angel* might be said to prime the mind of the listener to hear *pray* as well as *prey* when the [pre I] sound is pronounced by the actor. Here the countable text, like the modernized edition, must fall short, since both necessarily opt for a single possibility.

2.5 Benchmarking and disambiguating by frequency

Of course we are making some heroic assumptions here. Are the patterns of association regular enough to be truly discriminating in the ambiguous cases? Is the language of the BNC texts—prose fiction, mostly from the late twentieth century—similar enough to the language of sixteenth- and seventeenth-century plays so that information from one can usefully be applied to the other? We need a method of testing the accuracy of our disambiguations (and in fact to refine the method itself, so we know which of the various possible variations in methods work best).

One problem for WSD is that training examples generally need to be manually tagged, giving rise to a 'knowledge acquisition bottleneck'. Variant spelling disambiguation escapes this difficulty since the headwords come already correctly spelled in any

modern-spelling corpus, yielding as many training examples as there are occurrences of the headwords. We use Yarowsky's 'pseudo-words' technique and treat a modernized text as though it contained headword ambiguities (Yarowsky, 1993). In this way we already know the correct answer so can establish a success rate. For this we again used the thirty-seven plays in the public-domain Moby Shakespeare corpus, and added fifteen other Early Modern English plays available in modernized form. We examine each instance in these texts of the headwords we have classified as ambiguous and attempt to assign it to one of the two, or to one of the many, headwords in its group. If 'gentile' appears in a modernized play, we follow the presumption built into our variants system that this may be either an instance of *gentile* or an instance of *gentle*, and attempt to assign it to one or the other through its context. If we come up with *gentile*, it counts as a success; if it comes up as *gentle* it is not.

Figure 1 shows the degrees of success in assigning the pseudowords to the correct spelling using just one of the context words.

The 'cases' here are the 137,000 instances of the words our system defines as ambiguous in the thirty-seven Moby Shakespeare plays. A number of the cases (marked in the columns in lighter grey) were impossible to test, either because none of the context words occurred in the BNC corpus in those positions, so that there was no data to work on, or because there was too little context in the plays themselves, for example because they were isolated in a very short speech (we did not allow contexts to extend over a speech boundary). A second group (indicated by a texture pattern in that part of the column) was tested, but the results were excluded from consideration because one of the scores was beyond one of the thresholds we had established to weed out aberrations. An abnormally high score of this kind indicates that the method is performing unreliably in a particular case, as for example when a very rare word happens to appear once in the BNC near one of the candidate headwords and in the same slot in the context passage. In all the success rates derived from the modernized texts, we have to bear in mind that the corresponding rate may well be lower in the old-spelling texts, since the chance of finding context words is reduced by all the spelling variation.

The two right-hand columns of Fig. 1 show that choosing the headwords at random gives a correct result in fewer than a third of the cases (30%), while choosing the most common headword yields a surprisingly good result (success in 72% of the total of cases tested). The result from random choice reflects the average number of rival candidate headwords, just over three. (The system allows the user to set a limit for the number of candidates, acknowledging that with a large number the chances of success are small.) The success of the frequency method indicates that there is often an imbalance in that one candidate headword is much more common than the others. The greater the typical imbalance, the greater the success rate from this method.

Of the single-word slots, the most effective was the '−1' slot, the context word immediately preceding the ambiguous word. The next best was the '+1' slot. In fact the 'left-hand' slot always performed better than the corresponding 'right-hand' one. The OSTI researchers also found that the 'left-hand' context words were a little better as predictors of the target word for all but the most frequent words (Sinclair *et al.*, 2004, pp. xix, 48). In our results, single word slots beyond '−5' and beyond '+4' chose the correct headword at a rate below choosing at random.

Figure 1 gives the success rate of predictions using a single context word. We can combine the predictive power of the various word slots, and Fig. 2 shows the change in effectiveness as the 'window' of context is widened from '−1' to '+1' to '−9' to '+9'.

Next to the columns showing the results for various windows we show the results from choosing a headword at random and choosing the most common headword, as in Fig. 1. Performance improves markedly at first as more of the context words are used, then flattens off. There is a little improvement in expanding to eight context words from six, then no more after that. Given that going from six context words to eight is a significant extra demand on computer resources, and we wish the process to work 'on the fly', sometimes with very

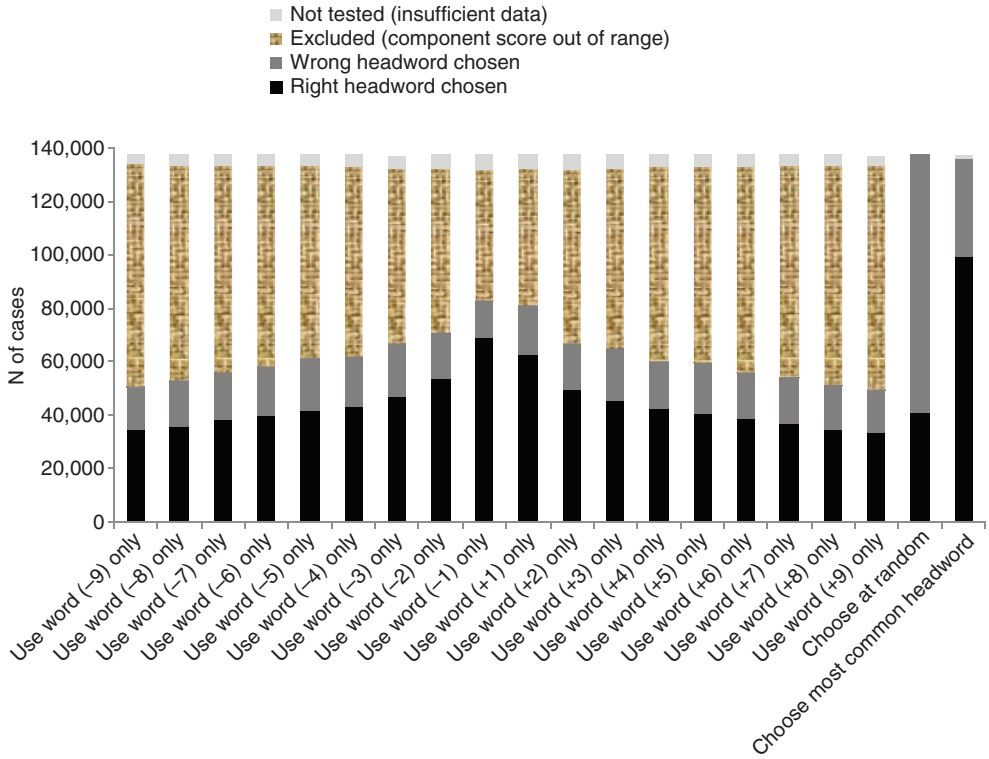

Fig. 1 Disambiguation results in a benchmarking set using one context word at a time

large corpora, we use the '−3' to '+3' window as our standard.

At this window size the performance of the system is much better than choosing at random, but the error rate is below what would be acceptable if we wish for 'accurate' word counts: of the cases involving words we would count, i.e. those not discarded, 19% of the ambiguous words are wrongly assigned. (Incidentally, this is similar to the error rate which results when we ask the machine simply to pick the most common out of the candidate headwords.) The method has power, but is not usable in practice. One error in five would make our word counts too unreliable.

The benchmarking system allows us to improve overall reliability by sorting the ambiguous cases from the beginning, leaving those where we are not confident of success unresolved, with their headwords removed from counting, while retaining a second group where we can anticipate a good result. To do this we can return to our test plays and look separately at each headword. What was the success rate, not on average, but for that particular word? The Intelligent Archive allows us to set a minimum success rate. The overall success rate will be much better than this. We use a standard minimum of 95%, so that the overall success rate is sure to be better than that.

Given the success rate achieved when we simply choose the most common headword, we can incorporate that method into the process. We can again set a minimum success rate in the benchmarking texts of 95%, and where both methods exceed the minimum, we can use the method with the superior success rate.

2.6 A composite system

The old-spelling task is to maximise compression, consolidating as many variants as possible into standard spellings, and at the same time minimise the

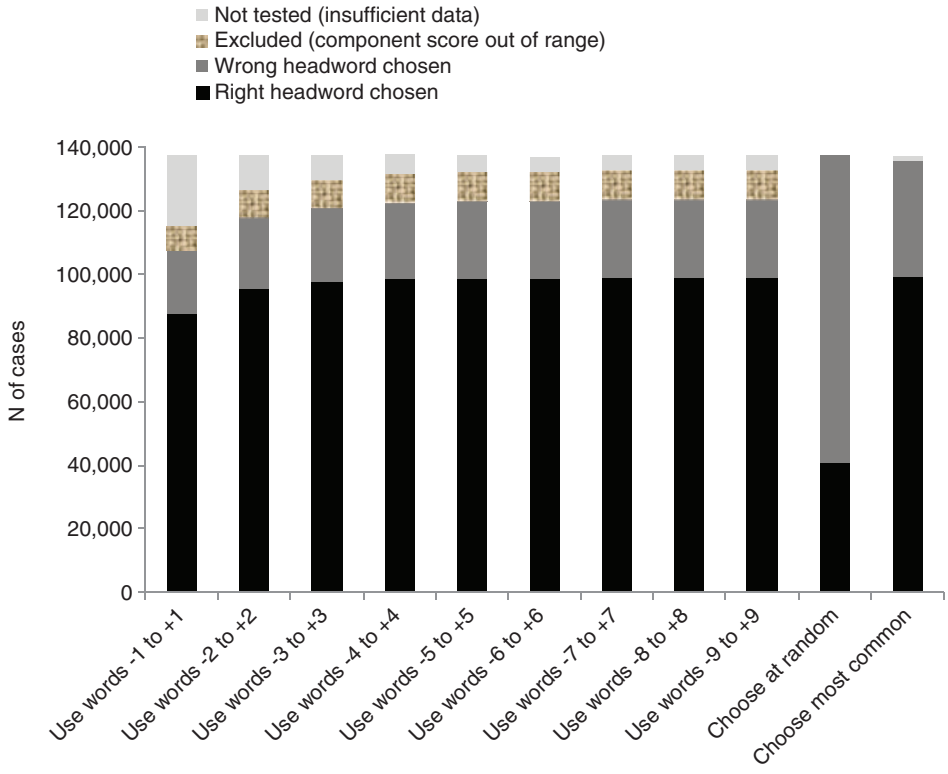

Fig. 2 Disambiguation results in a benchmarking set with various sizes of window

number of variants that cannot be resolved by the method and so must be discarded. As the variant lists grow, serving the aims of compression by capturing more and more variants of the various headwords, the chances of overlap between the list increases, and with it the number of ambiguous words.

Figure 3 shows how compression and the number of ambiguous words fluctuate with various methods. For this set of tests we used a corpus of 218 plays and poems, 3.6 million words in total.

The first column in Figure 3 shows raw counts, with no compression. The 'main count' total is 87,875, to which the 200 function words should be added, making 88,075 in all. Using the pairings which arise from comparing texts from the *Moby Shakespeare* with old-spelling ones makes for some compression, and also necessitates a tally of ambiguous words where the same spellings appear in more than one headword list (second column). Using the *OED* pairings makes for less compression and a larger collection of ambiguous words (third column). Combining the two sets of pairings (fourth column) makes for greater overall compression, and the largest collection so far of ambiguous words. If we add further standardization by way of spelling rules (treating 'u' and 'v' as interchangeable, and so on) we get still greater compression and a still larger count of ambiguous words, in fact these now outnumber the main count (fifth column). However, if we filter out some of the (by now very numerous) pairings in the headwords' lists, by excluding spellings which do not appear anywhere in this corpus, we decrease the proportion of ambiguous words to countable ones (sixth column). This step acknowledges that unchecked expansion of the lists, for example with Middle English variants from the *OED*, increases the chances of overlaps between lists to an unhelpful degree. (The system also allows for manual editing

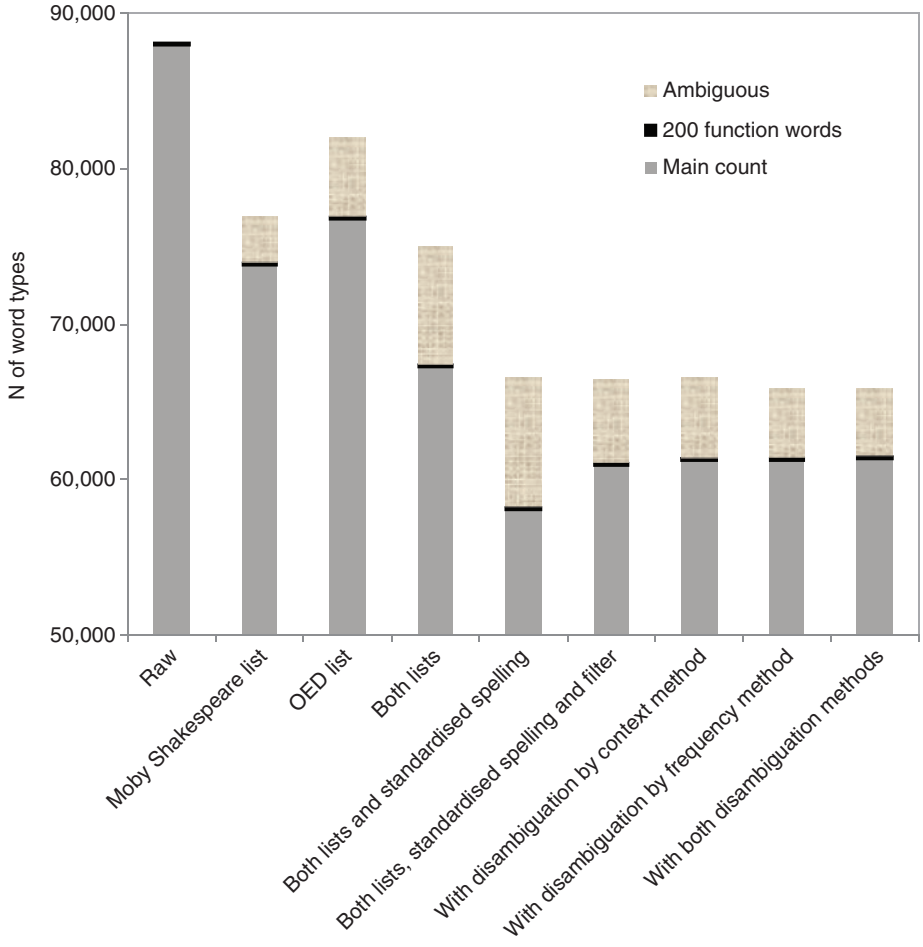

Fig. 3 Variation in counts of word types using various methods of compression and disambiguation

of the library of spelling equivalents, so as to eliminate the occasional anomaly thrown up by the text-matching procedure in particular.) The last steps are to eliminate some ambiguous words through disambiguation, either by the context method (seventh column), the frequency method, i.e. choosing the most common headword (eighth column) or a combination (ninth column). The context method used by itself returns 256 words from the ambiguous group to the main count, the frequency method by itself 290, and the two methods combined 350. These are small gains in the context of the overall numbers (each gain is only a small fraction of the 5,383 words in the ambiguous class prior to disambiguation, shown in the sixth column). This is explained partly by the threshold of 95% accuracy in the pseudowords benchmarking, but mostly by the complexity of the disambiguation task. The final count of usable words is 61,271 in the main count, plus the 200 function words, 61,471 words in all, to compare with the 88,075 words of the raw count.

3 Concluding remarks

3.1 The Intelligent Archive

The disambiguating system we describe here has been incorporated into a composite tool, the 'Intelligent Archive' or IA. The IA creates a corpus at need from an archive, counts word-forms

according to parameters chosen by the user, after the pre-processing of spelling variants already described, and then transforms the counts in various ways according to the experiment in hand. Its output is in the form of tables of counts and scores, which can be exported to a package like Excel or to SPSS for further analysis and for graphing. It is 'intelligent' to the extent that it can respond to requests for various selections and comparisons among its texts, can create and preserve various sub-corpora, and can assign most of the variant spellings it encounters in Early Modern English plays to the correct headwords. It has enough knowledge of the typical context for individual words in English generally to make good choices between headwords when a spelling is ambiguous. Presented with an unfamiliar play in old spelling from the 1580s or 1590s it can come up with counts for most of the words it contains, to a high degree of accuracy, modernizing the lexical words through rules, through background familiarity with the language of the time, and a knowledge of some of the underlying patterns of English.

3.2 Indeterminacy

A 'word' is an elusive entity. It may be a typographic or an orthographic form, or a dictionary headword. (In discussing frequencies it may also be a word type or a word token.) Old spelling complicates things further, though in another sense it helps reveal the inevitable degree of indeterminacy in all but the most rudimentary words data. To deal with this 'fuzziness', composite methods, work-arounds, and some manual intervention are needed. To arrive at the counts the Intelligent Archive in the end uses for statistical work, the software responds to the added indeterminacy of the early modern printed word with a series of transformations of the surface features of the printed versions, mixing direct interventions by the individual preparing the text for a limited set of features with on-the-fly processing by the software which draws on libraries of equivalents for the bulk of the work, and for the rest using two reference corpora, one to resolve ambiguities and the other to allot success rates to the disambiguation of individual words. This is a pragmatic, probabilistic endeavour. Computational methods need to respond to complicated, confused, and occasionally irreducibly indeterminate relationships.

Dealing with variant spelling is a major challenge for automated methods, and while they do make it possible to count words in old-spelling texts on a large scale, we have to accept that a percentage of error is built in, far beyond the occasional slip and the odd piece of guesswork in a text worked over case by case by an expert editor. To pursue quantitative linguistic work, the researcher must meet the difficulty wherever possible, build in as many safeguards as can be devised, and always, in working with results, bear in mind the probabilistic foundations of the base counts.

When we move away from the realms of copy-edited and proof-read enduring publications, to more ephemeral printed sources, and to speech mechanically or speedily transcribed, a surprising continuity between old-spelling and modern-era data emerges. *Pray* and *prey* were homophones and homographs in Early Modern English. They have remained homophones, but with standardized spelling they ceased to be homographs in a formal sense. There remains a tendency for the two forms to overlap, however.

3.3 Computational methods

One other conclusion from the present work is that the BNC fiction corpus, for all its 16 million words, is small. It is no surprise that 'Celestiall' as spelled in the 1604 *Hamlet* does not show up at all in this corpus of modern-spelling English, but even 'celestial' spelled the modern way appears just 25 times in 16 million running words. There is no instance of this word within three words of any instance of *pray* or *prey*. In fact, based on what happens in the BNC fiction corpus, a text collection would need to be over four and a half billion words in extent before one would expect an instance of *celestial* to occur by chance within three words either side of *pray*, and would need be close to ten billion words before one would expect to find an instance within three words either side of *prey*.[12] The zeros in Table 1 are another way of indicating how rare collocations between one word and another are for all but the commonest words. Even with a corpus twenty times bigger, like COCA, there are still surprising

gaps—surprising in the sense that a language user could formulate possible or even likely combinations that do not in fact occur.

Noam Chomsky argued long ago that this sort of limitation makes corpus study in general unprofitable for the understanding of language (Chomsky, 1957, p. 15).[13] He cited two nonsensical sentences, one of which was nevertheless grammatical, and suggested that since neither would occur in any conceivable corpus, corpus study could never distinguish the grammatical nature of one from the non-grammatical nature of the other. Pereira has responded by showing that a modern statistical model for collocation can assess the likelihood of any given passage occurring, based on predictive patterns in a corpus, and that one such model does indeed return a vastly higher level of probability for Chomsky's grammatical sentence than for his non-grammatical one (Pereira, 2000, p. 7). Corpus data will always be sparse in relation to the potential of a language system, but combining the information from multiple variables, and working on probabilities, can make for a useful performance, especially in a narrowly defined task like the disambiguation problem described above. Significant local regularities do emerge in the co-occurrence of words in the BNC fiction corpus, especially with the common function words, as the success rate of the context disambiguation method indicates. The error rate on the other hand shows that there is also a lot of variation. We can make some predictions on the basis of immediate context, but have to be prepared to be wrong quite often. Short of a gigantic corpus of a closely related language type, we need to discard some data, identified by a benchmarking process, to sharpen performance, and even then one must contend with a residual error rate. Language in general can be described as weakly probabilistic. Computational methods need to exploit the patterning that exists while guarding against its fallibility.

Working through the problems in standardizing old spelling texts automatically has provided vivid illustrations of the issues involved in applying computation to certain kinds of language. There is a constant tension between the goal of unequivocal definitive counts and what is possible with automated processing. On the one hand, the researcher has the use of a huge amount of data, captured within reasonable, and reasonably well-defined, limits of error; on the other, the researcher must deal with an uncomfortable indeterminacy, an amalgam of the errors which remain after multiple filtering and cross-checking procedures, and the (admittedly) much smaller number of ambiguities inherent in the irregular, 'unanchored' printed versions which are in the end our only connection with the theatre of Shakespeare and his contemporaries.

Notes

1 Randall McLeod warns that there may be an 'old typography' to reckon with as well as an 'old spelling', in that some expanded and contracted forms were clearly introduced by the compositor to help justify a line or to protect exposed kerns on pieces of type (McLeod, 1984, p. 81).

2 Bowers (1966, pp. 157–8) and Orgel (1996, pp. 56–7) give examples of significant problems in interpretation arising from this particular ambiguity.

3 Other examples of early modern English homonyms which are separate words in Modern English are 'loose' and 'lose', 'flower' and 'flour', 'draft' and 'draught', 'die' and 'dye', 'curtsy' and 'courtesy', 'cloths' and 'clothes' (Brook, 1976, p. 56).

4 Some words in the 200 function-word list discussed below are tagged manually for grammatical function.

5 Alistair Baron's Variant Detector tool, now in its second version ('VARD 2') implements a different solution to the problem, based on manually created mappings from variant to modern headword, letter replacement rules, and the Soundex phonetic matching algorithm. (In contrast, the software described in the present article relies mainly on a pre-installed library of mappings.)

6 For an example of a method using an individual classifier for each ambiguous word type, see Gliozzo *et al.* (2005).

7 Quotations from Shakespeare are from the Riverside edition. The square brackets mark editorial emendations.

8 The composite score for each headword is derived by this formula:

$$a_1/(b_1 * c_1 * d_1) + a_2/(b_2 * c_2 * d_2) + a_3/(b_3 * c_3 * d_3) + a_4/(b_4 * c_4 * d_4) + a_5/(b_5 * c_5 * d_5) + a_6/(b_6 * b_6 * d_6)$$

where $a_{1..6}$ are the counts for the word in the given position in the relevant part of the BNC, $b_{1..6}$ are the totals of instances of that word in the BNC, $c_{1..6}$ are the totals of instances of the headword in the BNC, and $d_{1..6}$ are the distances of the relevant word position from the target word, i.e. 1, 2, or 3. The sum of these component scores is multiplied by 1,000,000 for readability.

9 I am indebted to Harold Tarrant of the Hunter Bird Observers Club (NSW) for a further piece of evidence in favour of the *prey* reading. He suggests that the red kite (*milvus milvus*), which was common across England in Shakespeare's day, was well known for scavenging on refuse and may well be invoked by this passage. Elsewhere Shakespeare commonly refers to kites, always with a negative connotation (Armstrong, 1946, pp. 11–16).

10 COCA gives the source as Graubart, 2000.

11 COCA gives the title for the source story as 'The late Sam Boone: Bud Sparhawk' and categorises it as 'Analog Science Fiction & Fact'. Readers may notice that in this case 'on' does not immediately follow 'prey', so this sort of instance would not be counted in the '+1' position for the purposes of a calculation like that in Table 1. The formula we use is rigid. It is focused on counts of the same words in exactly the same positions, as in the particular juxtaposition 'pray on', rather than the more general case of 'pray' followed by 'on' whether immediately or after an adverb or other interruption. (Five other components are also taken into consideration in the overall score, of course.)

12 Sinclair has a comment on this aspect of collocations (Sinclair *et al.*, 2004, p. xxviii), as do Karov and Edelman, 1998, p. 42. As Sinclair points out, however, collocations clearly are not random (Sinclair *et al.*, 2004, p. xxii)

13 His distrust of corpus work remains, judging from a recent interview (Andor, 2004, pp. 96–100).

Acknowledgement

We are grateful to Peter Peterson for helpful comments on an earlier version of this paper.

References

Agirre, E. and Edmonds, P. (2006). Introduction. In Agirre, E. and Edmonds, P. (eds), *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer, pp. 1–28.

Andor, J. (2004). The Master and His Performance: An Interview with Noam Chomsky. *Intercultural Pragmatics*, **1**: 93–111.

Andrews, J. F. (1998). Site-Reading Shakespeare's Dramatic Scores. In Halio, J. L. and Hugh, R. (eds), *Shakespearean Illuminations: Essays in Honor of Marvin Rosenberg*. Newark, Delaware: University of Delaware Press, pp. 183–202.

Armstrong, E. A. (1946). *Shakespeare's Imagination: A Study in the Psychology of Association and Inspiration*. London: Lindsay Drummond.

Baron, A. (2008). *VARD 2*. http://www.comp.lancs.ac.uk/~barona/vard2/(accessed 30 July 2009).

Bowers, F. (1966). *On Editing Shakespeare*. Charlottesville, Virginia: University Press of Virginia.

Bowers, F. (1975). Old-Spelling Editions of Dramatic Texts. In *Essays in Bibliography, Text, and Editing*. Charlottesville: University Press of Virginia, pp. 289–95.

Brook, G. L. (1976). *The Language of Shakespeare*. London: Andre Deutsch.

Chadwyck-Healey Literature Online. (1996–2009). ProQuest LLC. http://lion.chadwyck.co.uk (accessed 30 July 2009).

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Crystal, D. (2008). *'Think on My Words': Exploring Shakespeare's Language*. Cambridge: Cambridge University Press.

Davies, M. (2008). *Corpus of Contemporary American English*. Brigham Young University. <http://www.americancorpus.org/> (accessed 23 December 2008).

de Grazia, M. (1990). Homonyms before and after Lexical Standardization. *Deutsche Shakespeare-Gesellschaft West Jahrbuch*, 143–56.

de Grazia, M. and Stallybrass, P. (1993). The Materiality of the Shakespearean Text. *Shakespeare Quarterly*, **14**: 255–83.

Early English Books Online. (2003–2009). ProQuest LLC. http://eebo.chadwyck.com/home (accessed 30 July 2009).

Egan, G. (2005). Impalpable hits: Indeterminacy in the searching of tagged Shakespearian texts. Shakespeare Association of America, Bermuda. Unpublished paper, available at <http://www.gabrielegan.com/publications/Egan2005a.htm> (accessed 12 January 2009).

Erne, L. (2003). *Shakespeare as Literary Dramatist*. Cambridge: Cambridge University Press.

Gliozzo, A, Giuliano, C, and Strapparava, C. (2005). Domain Kernels for Word Sense Disambiguation. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, June 2005.

Graubart, K. B. (2000). Weaving and the Construction of a Gender Division of Labor in Early Colonial Peru. *American Indian Quarterly*, **24**: 537–61.

Gurr, A. (2001). Other Accents: Some Problems in Identifying Elizabethan Pronunciation. *Early Modern Literary Studies*, **7.1**: 5.1–4.

Jackson, M. P. (2002). Determining Authorship: A New Technique. *Research Opportunities in Renaissance Drama*, **41**: 1–14.

Jackson, M. P. (2007). Is 'Hand D' of Sir Thomas More Shakespeare's? Thomas Bayes and the Elliott–Valenza Authorship Tests. *Early Modern Literary Studies*, **12**: 1.1–36.

Karov, Y. and Edelman, S. (1998). Similarity Based Sense Disambiguation. *Computational Linguistics*, **24**: 42–59.

LaDuke, W. (1999). *All Our Relations: Native Struggles for Land and Life*. Cambridge, Massachusetts: South End Press.

Leacock, C. and Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Fellbaum, C. (ed.), *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, pp. 265–83.

Leacock, C., Towell, G., and Vorhees, E. (1993). Corpus based statistical sense resolution. *Proceedings of the ARPA Workshop on Human Language Technology*, Plainsboro, New Jersey, pp. 260–5.

Madhu, S. and Lytle, D. W. (1965). A figure of merit technique for the resolution of non-grammatical ambiguity. *Mechanical Translation*, **8**: 9–13.

McLeod, R. (1984). Spellbound. In Shand, G. B. and Shady, R. C. (eds), *Play-Texts in Old Spelling*. New York: AMS Press, pp. 81–96.

Morton, R. (1984). How Many Revengers in *The Revenge Tragedy*? Archaic Spellings and the Modern Annotator. In Shand, G. B. and Shady, R. C. (eds), *Play-Texts in Old Spelling*. New York: AMS Press, pp. 113–22.

Orgel, S. (1996). *Impersonations: The performance of gender in Shakespeare's England*. Cambridge: Cambridge University Press.

Oxford English Dictionary. (1989). Second Edition. 20 vols. (Eds J. Simpson and E. Weiner). Oxford: Clarendon Press.

Pereira, F. (2000). Formal Grammar and Information Theory: Together Again? *Philosophical Transactions of the Royal Society*, **358**: 1239–53.

Rosenbaum, R. (2006). *The Shakespeare Wars: Clashing Scholars, Public Fiascoes, Palace Coups*. New York: Random House.

Salmon, V. (1986). The Spelling and Punctuation of Shakespeare's Time. In Wells, S. and Taylor, G. (eds), Shakespeare, W. *The Complete Works: Original-Spelling Edition*. Oxford: Clarendon Press, pp. xlii–lvi.

Schoenbaum, S. (1981). *William Shakespeare: Records and Images*. New York: Oxford University Press.

Shakespeare, W. (1997). *The Riverside Shakespeare*. Second Edition. Ed G. B. Evans *et al*. Boston: Houghton Mifflin, 1997.

Shakespeare, W. (n.d.). *Moby Shakespeare. The Complete Unabridged Works of Shakespeare*. Ed. G. Ward. Available at <http://icon.shef.ac.uk/Moby/>. Accessed 10/1/09.

Sinclair, J. M., Jones, S., and Daley, R. (2004). *English Collocation Studies; the OSTI Report*. Ed. Krishnamurthi, R. Continuum.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, D. A. (eds), *Machine Translation of Languages*. New York: John Wiley, pp. 15–23.

Wells, S. (1979). Modernizing Shakespeare's Spelling. In Wells, S. and Taylor, G. (eds), *Modernizing Shakespeare's Spelling: With Three Studies in the Text of 'Henry V'*. Oxford: Clarendon Press, pp. 3–36.

Yarowsky, D. (1993). One sense per collocation. *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, New Jersey, pp. 266–71.