Using the *OED* quotations database as a corpus – a linguistic appraisal

Sebastian Hoffmann
University of Zurich

1 Introduction

Over the past decades, the number of historical corpora available has steadily grown. Perhaps the best-known and most widely used is the Helsinki Corpus. (See Kytö 1996[1991] for a description of the corpus and Rissanen et al. 1993 for a range of possible applications.) Other historical corpora include ARCHER (A Representative Corpus of Historical English Registers), the Corpus of Early English Correspondence (CEEC), the Innsbruck Computer Archive of Machine-Readable English Texts (ICAMET), the Lampeter Corpus of Early Modern English Tracts, and the Zurich English Newspaper Corpus (ZEN), to name just a few (cf. Biber et al. 1994; Fries 1994; Schmied 1994; Keränen 1998; Markus 1999a). However, given their relatively small size, these historical corpora are unfortunately only of limited value for the study of less frequent features of the English language. The Helsinki Corpus, for instance, spans almost a thousand years (ca. 750 to 1700) but contains only 1.57 million words. Even for the period of Late Modern English, suitable corpus data is not in great abundance. For example, although ARCHER covers a smaller time-span from 1650 to 1990 and offers detailed categorization by register, its overall size of less than two million words still results in many of the same limitations as the Helsinki Corpus.[1]

For the study of less frequent features, the researcher therefore has to make use of alternative – albeit potentially less reliable – sources of data. One of the options available is the *Oxford English Dictionary* (*OED*) with its large quotations database, covering more than a thousand years of English usage. This database, which is considerably larger than any of the historical corpora mentioned above, has been successfully employed to trace both lexical and grammatical changes over time (e.g. Jucker 1994; Fischer 1997; Markus 1999b and 2001; Mair 2001). However, to my knowledge there is no detailed appraisal of the *OED* quotations database as a tool for linguistic research. The present paper is intended to fill this gap.

2 The OED *quotations database*

The *Oxford English Dictionary* is generally considered to be the world's most comprehensive dictionary of the English language. Its compilation was started in the second half of the 19th century but it was not until the year 1928 that the first edition finally reached completion.[2] In 1989, the second edition, which incorporated the four-volume supplement issued between 1972 and 1986, appeared in twenty volumes. In 1987, a CD-ROM version of the first edition was released, which gave the user unprecedented access to a wealth of information about the English language. The second edition of the *OED* became available on CD-ROM in 1992, thereby extending the electronically accessible data to cover the complete history of the English language from its earliest extant texts until well into the second half of the twentieth century. (See Jucker 1994 for a review of the CD-ROM from a linguist's point of view and Johansson 1996 for an overview of possible applications.)[3]

The makers of the *OED* pursued an ambitious aim: not only was their dictionary intended to contain every word ever used in the English language, but also to document the "development of form and meaning" of each word illustrated with "a series of quotations ranging from the first known occurrence of [a] word to the latest, or down to the present day; the word being thus made to exhibit its own history and meaning" (Murray 1888: vi). In total, more than five million quotations were collected for this purpose by countless volunteers and over 1.8 million of these quotations were used in the first edition of the *OED*. An additional 600,000 quotations were then added for the release of the second edition.[4] Using the program provided with the CD-ROM, this large database of over 2.4 million quotations can be searched for individual lexical items or phrases and thereby provides computerized access to samples of the English language spanning a period of more than 1,000 years.

The main question to be answered in the present context is whether the *OED* quotations can be employed for a meaningful linguistic analysis of earlier stages of the English language which goes beyond purely qualitative description. Asked even more succinctly: can the *OED* quotations be used as a corpus? To answer this question, a number of aspects require attention. The following four points will be treated in more detail:

- Selection criteria for the quotations
- Representativeness and balance of the quotations
- Reliability of the data format
- Quantification of the results

3 Selection criteria for the quotations

First of all, consider the following standard definition of a corpus:

> A *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. A *computer corpus* is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance. (Sinclair 1996; emphasis in the original)

The principal stumbling block for counting the *OED* quotations as a corpus is posed by the selection criteria. Although the individual quotations were indeed selected according to explicit linguistic criteria, their main purpose is to exemplify the meaning and use of a particular word with a minimal amount of context. Thus, the material was obviously not collected with a view to creating a representative sample of the language for a particular period of time. As Sinclair (1996) writes, it is important to keep clear the distinction between true corpora and mere collections of citations:

> Citations are individual instances of words in use and collections of these also have no claims to be corpora. The precise conditions for a valid sample size for a corpus are indeed under discussion ... but no-one concerned seriously with corpora has attempted to gather a collection of citations and announce it as a corpus. What has happened is that owners of previously-gathered citation collections have tried to use them as a bridge between traditional practice – particularly in lexicography – and corpus-based work.
>
> It is unhelpful to confuse categories in this way, and important to assert minimal criteria for use of the word ‘corpus’. (Sinclair 1996)

I fully agree with Sinclair on the importance and usefulness of this distinction. I nevertheless believe that the *OED* quotations database should not be completely dismissed as a source of quantitative data. The crucial question is whether it can be assumed that the repeated citation of a particular word correlates with a corresponding level of currency. To answer this question, a distinction must be made between all of the quotations for a given headword and the quotations database in its entirety. In the first case, the quotations were specifically selected to display the whole range of possible uses for a single word. These uses may include idiosyncrasies and relatively obscure variants. A correlation with actual currency in such circumstances is highly unlikely. However, apart from the headword, the quotations also contain other linguistic material which helps to place the headword in the proper context. These additional words, which typically constitute a large part of the quotation, occur in a much more unsystematic way. It is fair to assume that they simply reflect language use at the time of its writing. The sum of all of the *OED* quotations is therefore to a large extent made up of naturally occurring language. As a consequence, a researcher who uses the *OED* CD-ROM to search through the complete quotations database rather than only the set of quotations belonging to a particular headword should indeed be able to use this data for both qualitative and quantitative research. However, a prerequisite for such an application is complete awareness of the merits and limitations of the *OED* quotations. Some of these aspects will be presented in the following sections.

4 Representativeness and balance of the quotations

It is a fundamental property of language corpora that they represent collections of actual language use as produced by a cross-section of the speakers (and/or writers) using a particular language or variety. This is clearly also the case for the *OED* quotations. First, virtually all of the citations are true quotations; i.e. they are not constructed examples. The main exceptions are explained in Berg (1991):

> [I]n the first edition, when no examples were found of contemporary usage, illustrations were occasionally ‘made up’. These are introduced by the word ‘Mod.’ for ‘modern’ and normally appear as the last quotation in a paragraph without a date. In a few instances, portions of nursery rhymes or proverbs are quoted as examples of usage with no actual source, other than prefatory wording such as ‘Nursery Rime’, ‘Mod. Prov.’ (Modern Proverb). (Berg 1991: 36)[5]

Second, the range of sources for the quotations is extremely varied. It goes far beyond the 19th century practice of including only the works of “the best writers” (cf. Willinsky 1994, Chapter 2). The editors emphasized description and historical completeness rather than being arbiters of style. As a consequence, even the first edition of the *OED* contained a considerable number of quotations from periodicals (e.g. *London Gazette* and *Sporting Magazine*), and such non-literary works as the *Encyclopaedia Britannica* and the *Practical Dictionary of Mechanics* feature among the top twenty books cited.[6] In the second edition, the proportion of non-literary texts is even higher. It is also worth noting that the bibliographical information for the quotations listed in the second edition of the *OED* covers 143 finely printed pages. In terms of content, the *OED* quotations database is thus much more varied than any of the historical corpora mentioned in the introduction.

Despite the great variety in the contents of the *OED* quotations, it would be a mistake to regard them as a reasonably balanced representation of the English language. The proportions of the different types of quotations (e.g. fiction vs. non-fiction) clearly do not constitute a true mirror of the actual language use – or type of language exposure – of a particular period or area. It certainly was not one of the declared aims of the compilers to establish a comprehensive cross-section of all of the typical text domains. As a consequence, certain authors are overrepresented (e.g. Shakespeare's works contribute almost 33,000 quotations to the first edition of the *OED*), while other sources, such as working-class newspapers of the 19th century, are hardly featured at all.[7] Moreover, the number and range of texts included from English-speaking nations other than England are relatively limited, even today.[8] Seen from a corpus linguist's perspective, such considerations of content and composition should be kept in mind when reviewing the range of possible applications of using the *OED* quotations as a database.

5 Reliability of the data format

A further point in the linguistic evaluation of the *OED* quotations concerns the reliability of the data format. It must be remembered that the principal purpose of a quotation is the illustration of the meaning and use of a particular word. In many cases, this purpose is fulfilled even if parts of the quotation, such as subordinate clauses, are deleted. In the *OED*, such deleted elements in the quotation are represented by two dots (..). A typical example of this marked deletion can be seen when the sentences in (1) and (2) are compared:

(1) Now by my honor, *my life, my troth,* I will appeach the Villaine.

(2) Now by mine honor..I will appeach the Villaine. (*OED*, 1593 Shakespeare *Richard II*, v. ii. 79; *appeach* v.)

Example (1) is a sentence taken from Shakespeare's *Richard II* and (2) shows the corresponding illustrative quotation cited in the entry for the verb *appeach*. The deleted elements are marked by italics in sentence (1). From the point of view of the researcher, such deletions will – at least for most types of linguistic investigation – only have a marginal effect on the results. For example, although lexical material was indeed deleted from the original, the overall sentence structure has clearly not been affected.

However, other types of deletion can be found in the quotations database which may have a more troublesome impact on the outcome of the linguistic results. As a point in case, consider a sentence taken from Shakespeare's *Much Ado About Nothing* and its corresponding *OED* quotation, found in the entry for the noun *bird's-nest*:

(3) *The flatte transgression of* a Schoole-boy, *who being* ouer-ioyed with finding a birds nest, *shewes it his companion, and he steales it*. (*Much Ado About Nothing*, 1623 folio)

(4) A Schoole-boy..ouerioyed with finding a birds nest. (*OED*, 1599 Shakes. *Much Ado* ii. i. 229; *bird's-nest* n.)

Despite the punctuation, it is immediately apparent that the quotation in example (4) does not form a complete sentence. Furthermore, for a linguist interested in the historical development of certain grammatical structures, the difference between (3) and (4) may indeed have serious implications. For example, the postmodification by the non-finite *-ing* clause is completely changed in the *OED* quotation.

Although the *User's Guide to the Oxford English Dictionary* gives the impression that deletions are relatively rare (it uses the word “occasionally”), a large number of quotations do in fact contain marked deletions (cf. Berg 1991: 40). Table 1 presents information about the proportion of quotations which have undergone shortening in a number of selected years:

Table 1: The proportion of shortened *OED* quotations in selected years[9]

Year(s)	n quotations	n shortened quotations	%
< 999	9,462	1,763	*18.6*
1001–1099	1,503	232	*15.4*
1151–1199	4,131	762	*18.4*
1251–1280	1,872	250	*13.4*
1351–1365	2,203	347	*15.8*
1451–1458	2,029	498	*24.5*
1551	1,986	532	*26.8*
1651	4,764	982	*20.6*
1751	2,966	716	*24.0*
1851	8,100	1,626	*20.0*
1951	5,596	1,044	*18.7*

When dealing with sources dating from before the 15th century, the editors were presumably reluctant to shorten quotations, lest the modern reader's comprehension of them be jeopardized; between one fifth and one quarter of all quotations dating from later sources, however, were shortened. The actual number of deletions is even higher since approximately 15–20 per cent of the shortened quotations contain more than one marker of deletion. Although, admittedly, sentences where the shortening of quotations resulted in such radical changes as documented in examples (3) and (4) are unusual, the implications for a linguistic analysis should still not be disregarded.

A second point in connection with the reliability of the data format concerns the fact that deletion was apparently not always consistently marked. This becomes obvious when sentences (5) and (6) are compared:

(5) That ye neuer by way of curiosite be besy to attempte ony persone therin. (*OED*, 1526 Pilgr. Perf. (W. de W. 1531) 2; *curiosity*)

(6) I requyre you..that..ye neuer by way of curiosite be besy to attempte ony persone therin. (*OED*, 1526 Pilgr. Perf. (W. de W. 1531) 2; *way* n.)

The first of these quotations is found in the entry for the noun *curiosity,* whereas sentence (6), using the same source, illustrates the noun *way.* Not only do the sentence boundaries differ in these two examples, but there is also a marked deletion between *that* and *ye* in sentence (6), which is absent from the quotation shown in (5). Such examples appear to be quite rare. However, having said this, such instances are only detectable when the same quotation is used in (at least) two different formats. The actual number of unmarked deletions therefore remains unknown.

It is difficult to assess the full implications of edited *OED* quotations for the study of linguistic phenomena. However, particularly in the case of studies that do not investigate larger constructions, whose features span across clause boundaries, it seems reasonable to assume that the number of potential distortions is fairly limited.

6 Quantification of the results

The final point I would like to raise concerns the use of *OED* quotations for the quantification of linguistic trends. On the basis of a computerized diachronic source such as the Helsinki Corpus, normalized frequency counts for individual features are easily obtained. Such frequency information then makes it possible to compare the use of a particular feature or construction across different periods of time or text domains. In the case of the *OED* quotations database, such information is somewhat more difficult to retrieve. Absolute frequency counts for lexical items or phrases can be easily obtained by exporting the result of a search over the quotations text to a file and sorting this output by the year of quotation. In order to normalize this data, information about the number of quotations and the length of quotations is required. Figure 1 gives an overview of these two variables for the period between 1000 and 1980.

Not surprisingly, the number of quotations is very low for the first few centuries and it is not until the 14th century that the threshold of 1,000 quotations per year is crossed. The first peak is found at the beginning of the 17th century and is followed by a considerable drop in the 18th century. From about 1800 onwards, the number of quotations increases dramatically and reaches a peak with over 10,000 quotations per year when the first edition was being compiled. The beginning of the 20th century, however, is again somewhat underrepresented.

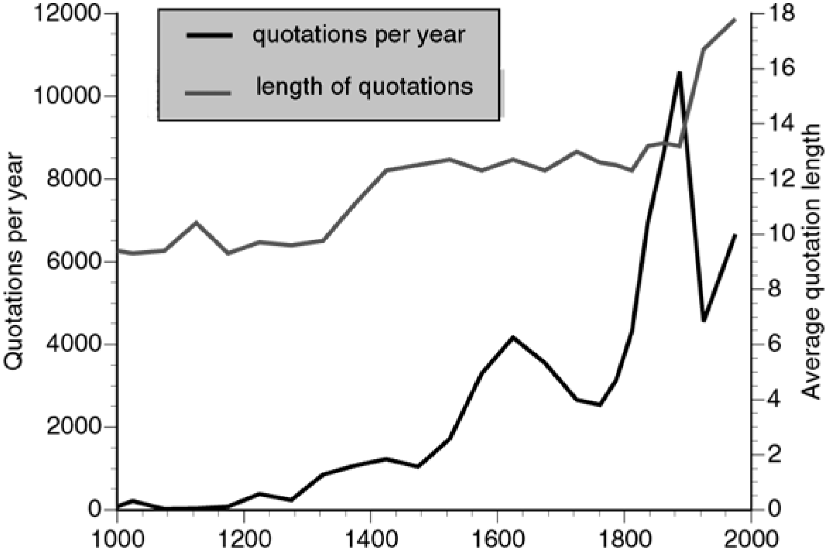

Figure 1: The number of quotations in the OED per year (averages over periods of 25–50 years) and their average word length

The second important variable is the average length of the quotations. This variable proves to be fairly constant, particularly for the time between 1450 and the end of the 19th century (approximately 13 words per quotation). Interestingly, the quotations added in the second edition of the *OED* are significantly longer (more than 16 words per quotation). It thus appears that the 20th century editors felt that more context was needed in order to demonstrate the meaning and usage of a word.

Until now, I have referred to the size of the quotations database by number of quotations. The data shown in Figure 1 can be used to approximately calculate the total number of words in the database: I estimate that the quotations database of the second edition of the *OED* comprises a total of 33–35 million words.

The data presented in Figure 1 clearly shows that absolute frequency counts retrieved from the quotations database of the *OED* require careful evaluation before they can be interpreted as indicative of a linguistic change in progress. Given the dramatic rise in the number of quotations during the 19th century, for example, a corresponding rise in the absolute frequency of a particular phenomenon in these quotations is of course to be expected.

On the basis of the information contained in Figure 1, normalized frequency counts can indeed easily be calculated. However, given the nature and purpose of the quotations database (cf. Sections 3–5), the researcher may perhaps feel that these counts should only be interpreted as an indication of possible trends. When analyzing diachronic patterns identified in the *OED* quotations database, it might therefore be considered more valid and reliable to limit the discussion of the data to general tendencies.

7 Conclusion

Although the *OED* quotations database is not a completely balanced and representative corpus, it can nevertheless provide the linguist with a wealth of useful information. The data it contains chiefly represents naturally occurring language, and the time-span covered is unmatched by any other source of computerized data. Even though over 20 per cent of all its quotations have been shortened, the large majority of these deletions is unlikely to distort the results of many diachronic studies of linguistic features. Given the nature of the data, normalized frequency counts might suggest an inappropriate level of precision, but tendencies in the development over time can nevertheless be expressed in quantitative terms.

The meaningful interpretation of corpus-based results is always highly dependent on the researcher’s awareness of the advantages and shortcomings of the dataset used. This statement is of course also true for linguistic investigations carried out with the help of the *OED* quotations database. It is hoped that the present paper has contributed useful information for historical linguists whose research requires a larger set of data than is provided by the corpora currently available.

Notes

1. In addition, some of the aforementioned corpora are restricted to special genres (e.g. the CEEC), which makes a meaningful comparison with other diachronic corpora or present-day data highly problematic.
2. The first fascicle of the *OED* – then still called *The New English Dictionary* – appeared in 1884. At the time of completion in 1928, the dictionary covered ten volumes and contained a total of 214,165 entries. This first edition was reprinted in 1933 in twelve volumes under the new name *Oxford*

English Dictionary. The year 1933 also saw the completion of a first supplement to the *OED*, which added a further 28,722 entries and which was appended to the first complete republication.

3. The CD-ROM version of the first edition was taken off the market after only three years. See Markus (1999b) for a brief discussion of the advantages of this first edition for historical linguists. The CD-ROM containing the second edition of the *OED* is currently available in its third version. It offers updated search facilities but is unfortunately restricted to users working with a Windows operating system. For further information, see http://www.oed.com.
4. For exact figures concerning the number of quotations in the individual components which were used to form the second edition of the *OED* (i.e. the 1933 supplement, the four-volume supplement of 1972–1986, and the New English Word series), please see Willinsky (1994: 209).
5. A researcher making use of the *OED* quotations may of course choose to exclude such constructed quotations from his or her analysis.
6. The top twenty periodicals by citation in the *OED* (first edition) covered about 80,000 quotations (4.4% of the total of 1,827,306 quotations). Willinsky (1994: 209ff.) presents a number of tables containing detailed statistics on the composition of the *OED* quotations database.
7. As Willinsky (1994: 11) notes, “[t]he OED will always represent something of the times in which it is being edited, as it absorbs common concerns about the state of the language”. He devotes a whole chapter entitled “The Sense of Omission” to the many other limitations in the range of illustrative *OED* quotations (see Willinsky 1994: 176ff.). For further insightful comments on the sources of the *OED* quotations, see also Brewer (2000).
8. In addition, regional differences are difficult to capture since the bibliographical information given for individual quotations is often relatively sparse. In her study of the influence of American English on Australian and British English, Peters (2001: 306) makes extensive use of the *OED* quotations database. She describes the difficulties she faced because “[t]he sources to which [the quotations] were attributed were often enigmatic, just a name and initial, plus date: S.E. White 1901. The *OED* bibliography adds only title of publication but not publisher, so it remains unclear what regional variety they represent.”
9. Since a clean text version of the full quotations database is not available, only data for selected years can be presented. In order to calculate the figures presented in Table 1, all of the quotations in a particular year (or a

sequence of years when the number of quotations was too low) first had to be exported to a text file.

10. The slightly lower figure for the year 1951 is not a coincidence. All of the other years which I checked from the middle of the twentieth century also showed similarly low percentages. This appears to be a reflection of different editorial practices for the second edition of the *OED*.
11. A second possible explanation is that the funds available to the editors were considerably increased.
12. In the introduction to the second edition of the *OED*, the total number of printed words is given as 59 million. This figure of course includes all definitions, etymological information, spelling variants etc.

References

Berg, Donna Lee. 1991. *A User's Guide to the Oxford English Dictionary.* Oxford: Oxford University Press.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its Challenges: Compiling and Exploring A Representative Corpus of Historical English Registers. In U. Fries, G. Tottie and P. Schneider (eds.). *Creating and Using English Language Corpora,* 1–14. Amsterdam: Rodopi.

Brewer, Charlotte. 2000. *OED* Sources. In L. Mugglestone (ed.). *Lexicography and the OED*. Pioneers in the Untrodden Forest, 40–58. Oxford: Oxford University Press.

Fischer, Andreas. 1997. The *Oxford English Dictionary* on CD-ROM as a Historical corpus: *To wed* and *to marry* Revisited. In U. Fries, V. Müller and P. Schneider (eds.). *From Ælfric to The New York Times: Studies in English Corpus Linguistics*, 161–72. Amsterdam: Rodopi.

Fries, Udo. 1994. ZEN – Zurich English Newspaper Corpus. In M. Kytö, M. Rissanen and S. Wright (eds.). *Corpora across the Centuries*, 17–18. Amsterdam: Rodopi.

Johansson, Stig. 1996. Introducing the Machine-Readable *Oxford English Dictionary. Image* 3/1: 19–38.

Jucker, Andreas H. 1994. New Dimensions in Vocabulary Studies: Review Article of the *Oxford English Dictionary* (2nd edition) on CD-ROM. *Literary and Linguistic Computing* 9/2: 149–154.

Keränen, Jukka. 1998. The Corpus of Early English Correspondence: Progress Report. In A. Renouf (ed.). *Explorations in Corpus Linguistics. Language and Computers: Studies in Practical Linguistics 23*, 29–37. Amsterdam: Rodopi.

Kytö, Merja (comp.) 1996[1991]. *Manual to the Diachronic Part of The Helsinki Corpus of English Texts*. Coding Conventions and Source Texts. 3rd edition. Helsinki: Department of English, University of Helsinki.

Mair, Christian. 2001. Early or Late Origin for *Begin* + V-*ing*? Using the *OED* on CD-ROM to Settle a Dispute between Visser and Jespersen. *Anglia* 119: 606–610.

Markus, Manfred. 1999a. Manual of *ICAMET* (*Innsbruck Computer-Archive of Machine-Readable English Texts*). Innsbrucker Beiträge zur Kulturwissenschaft, Anglistische Reihe, Vol. 7. Innsbruck: Leopold-Franzens-Universität Innsbruck, Institut für Anglistik.

Markus, Manfred. 1999b. English Historical Lexicology in the Age of Electronic Reproduction: Some Suggestions. In W. Falkner and H-J. Schmid (eds.). *Words, Lexemes, Concepts – Approaches to the Lexicon*, 365–378. Tübingen: Gunter Narr.

Markus, Manfred. 2001. Linguistic Commercialism in and around the *Paston* and *Cely Letters*. An *OED* and Corpus-Based Approach. *Journal of English Linguistics* 29/2: 162–178.

Murray, James A. H. 1888. *A New English Dictionary on Historical Principles*. Vol. 1. Oxford: Clarendon Press.

The Oxford English Dictionary. 1st ed., 12 vols. 1933. James A.H. Murray, Henry Bradley, W.A. Craigie & C.T. Onions (eds.). Oxford: Oxford University Press.

The Oxford English Dictionary. 2nd ed., 20 vols. 1989. Prepared by J.A. Simpson and E.S.C. Weiner. Oxford: Clarendon Press.

Peters, Pam. 2001. Varietal Effects. The Influence of American English on Australian and British English. In B. Moore (ed.). *Who's Centric Now? The Present State of Post-Colonial Englishes,* 297–309. Oxford: Oxford University Press.

Rissanen, Matti, Merja Kytö and Minna Palander-Collin (eds.) 1993. *Early English in the Computer Age: Explorations through the Helsinki Corpus*. Berlin: Mouton.

Schmied, Josef. 1994. The Lampeter Corpus of Early Modern English Tracts. In M. Kytö, M. Rissanen and S. Wright (eds.). *Corpora Across the Centuries*, 81–89. Amsterdam: Rodopi.

Sinclair, John McH. 1996. *EAGLES. Preliminary Recommendations on Corpus Typology*. http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html. (Accessed 16.11.2003).

Willinsky, John. 1994. *Empire of Words. The Reign of the OED*. Princeton, NJ: Princeton University Press.