Cawdrey's A Table Alphabeticall: A Quantitative Approach

Rosamund Moon
Department of English, University of Birmingham
Birmingham B15 2TT, Great Britain

Abstract
This paper investigates the headwords in Cawdrey's *A Table Alphabeticall*, an English monolingual dictionary first published in 1604, and covering 'hard words'. It explores ways of estimating how different the headword selection is from that of current dictionaries, and considers what 'hard words' may actually be in current terms.

1 Introduction

Robert Cawdrey's *A Table Alphabeticall* [CTA] first appeared exactly four hundred years ago in 1604. It is generally said to be the first monolingual dictionary of English, and contains about 2500 headwords, with definitions. These headwords are 'hard words', words which Cawdrey expected to be unfamiliar in some way to his readers. Since *CTA* is such an early lexicographical work and covers only part of the lexicon, the vocabulary that it represented is inevitably very different from that of modern standard dictionaries. This paper sets out to assess just how different it is.

2 Background

CTA is the first English dictionary in the sense that it is the first lexical reference book which contains a list of English words where all those words are explained in English. However, it is clearly an inheritor of lexicographical traditions that were already long established in England, as in continental Europe. Stein (1985) provides extensive documentation and discussion of many of *CTA*'s English antecedents, from early bilingual glossaries through bilingual dictionaries of the 15th and 16th centuries: see also Schäfer (1989) for discussion of dictionaries of this period, and Green (1996: 147ff and passim) for a general historical overview. Other antecedents include 16th century pedagogical works, with advice and information on grammar, writing, and spelling: the last of these typically included or comprised word lists. Of these antecedents, two are particularly important as direct sources for *CTA*. One is Edmund Coote's *The English School-maister* (1596): this included instructional data on spelling, grammar, and religion, as well as a list of around 1400 English vocabulary items, mostly 'hard words' and mostly glossed. Starnes and Noyes (1991: 13ff) estimate that 87% of Coote's proto-dictionary was incorporated into *CTA*, and 40% of *CTA* is taken from Coote. The other is Thomas Thomas's *Dictionarium Linguae Latinae et Anglicanae*, a unidirectional Latin-English dictionary published c. 1587 with nearly 40000 headwords: see Stein (1985: 312ff) for discussion. Starnes and Noyes (1991: 15ff) estimate that a further 40% of *CTA*'s headwords is drawn from Thomas: in addition, many items taken

from Coote are amplified with material from Thomas. There is further discussion of sources in Siemens (1994).

The focus on 'hard words', which characterizes Coote, *CTA*, and English dictionaries of the following decades, can be traced back to traditions of early glossaries. Since hard words were, loosely, words formed from Latin and Greek roots which had been identified in some way as outside the central vocabulary of English, they corresponded to that central vocabulary in much the same way as did words of other languages. By 1604, of course, early glossaries had already evolved into bilingual dictionaries, which were both sophisticated and broad in scope. Thomas, for example, included information on grammar and usage, and dealt with common function words as well as content words. One early monolingual word list, the list of 8000 undefined items in Mulcaster's *Elementarie* (1582), similarly included function words and other central items along with more morphologically complex words. However, the perceived need in *CTA* and Coote was the provision of information about hard words, not an inventory of English vocabulary.

CTA's purpose was clear, and Cawdrey set it out on the title page (orthography here, apart from the title, has been modernized):

> A Table Alphabeticall, containing and teaching the true writing, and understanding of hard usual English words, borrowed from the Hebrew, Greek, Latin, or French, &c. With the interpretation thereof by plain English words, gathered for the benefit and help of ladies, gentlewomen, or any other unskilful persons. Whereby they may the more easily and better understand many hard English words, which they shall hear or read in scriptures, sermons, or elsewhere, and also be made to use the same aptly themselves. *Legere, et non intelligere, neglegere est.* As good not read, as not to understand.

This identified the target demographic as those who were lacking educationally in some respect, notwithstanding the norms for their general social class – specifically, women. Further, it identified the function of the dictionary as primarily pedagogical, aiming at both decoding and encoding: this pedagogical function is important. The subtext is that these words had now become both accepted and acceptable, part of normal educated vocabulary, 'hard usual words' to be understood and used 'aptly'. This appears to contrast with the ideological stance taken by many earlier English scholars, who condemned morphologically alien, complex words as inkhorn terms, unnecessary additions to the lexicon. In fact, Cawdrey himself went on to discuss and condemn inkhorn terms in the preface to *CTA*, distinguishing them from his selected headwords, which he manifestly regarded as neither abnormal nor to be avoided. See Hayashi (1978: 31ff) for a discussion of inkhorn terms and hard words, and the relationship between them.

2.1 A Note on the Text

Observations in this paper draw on a version of *CTA* which was reproduced from a manuscript in the Bodleian library, Oxford, and is also the basis for the electronic version created by Ian Lancashire at the University of Toronto web site. There are a number of imperfections in this text, including printers' errors. In particular, there is no text in this

edition for letters *K* and *W/X/Y*, although later editions of *CTA* include entries in those ranges. The absence of header letters at the beginning of the ranges *L* and *Z* and the fact that Coote included four words in *w-* suggests that text had been omitted by mistake.

Osselton (1989: 165ff) comments on printers' errors in *CTA* in his discussion of its alphabetical sequencing. As he points out, there are a number of deviations: headwords in *CTA* are not strictly alphabetical, even taking into account contemporary practices where words were not necessarily ordered beyond their first two letters. Early word lists and glossaries had generally been thematic: see Stein (1986: 219f) for further comments on sequencing in early English dictionaries. In addition, orthographic conventions of the time meant that letters *i* and *j*, and *u* and *v* were interchangeable, depending on word position: when initial, words were listed in single alphabetical ranges, *i/j* and *u/v*.

The following is largely based on *CTA* in its original orthography. However, I also normalized forms and spellings in order to facilitate automatic comparisons with distributions in current English.[1]

3 Headwords and Currency

Approximately 50% of *CTA*'s headwords survive unchanged into modern English: that is, unchanged in either meaning or spelling. Another 40% have changed orthographically, although these changes are mostly minor. Such changes typically involve doubled consonants, the representation of /k/ and unstressed or long vowels, and different conventions with the use of *i/j* and *u/v*: hence *acquitall/acquittal, academicke/academic, adiacint/adjacent, acheeue/achieve*, and *analogie/analogy*. A few words have changed morphologically: *destinated/destined, insociable/unsociable, patheticall/pathetic*, and *perspicacy/perspicacity*. There are a few cases of erroneous forms and printing errors: *falcinate* for *fascinate*, *suggect* for *suggest*.

Around one in eight of those headwords in *CTA* which are still extant are defined in senses which are either now technical or else not the current primary sense. For example, the sense given for *probleme/problem* is 'proposition, or sentence in manner of a question', and for *traffique/traffic*, 'bargayning'. Other headwords in *CTA* are either now obsolete, or recorded in senses which are obsolete: for example, *frigifie* 'coole, make cold', and *temperature* 'temperatenes, meane or due proportion'.

Headwords in *CTA* which have survived have widely-varying frequencies in current English, as shown in the following table. This takes into account the grammatical class of items as defined in *CTA*, but not their meanings, and maps them onto frequencies in the Bank of English corpus: 450 million words of current English.[2]

no. of headwords	*frequency range*
85	over 100 tokens per million
531	10-100 tokens per million
32	1-10 tokens per million
482	1 token per 1-10 million
175	1 token per 10-100 million

Table 1: Extant headwords and their frequencies in current English

Other headwords have frequencies below this point or are not attested in the corpus. Of the twenty *CTA* headwords which are highest in frequency in current English, half were already extant in those senses in the 14th or 15th centuries: by including them, Cawdrey further validated them and demonstrated their long-term success as words. A few of these twenty headwords, however, were relatively new in 1604: that is, not more than 25 years old. These include *national*, and, in the senses given in *CTA*, *centre*, *expect*, *real*, and *social*.

> centre, middest of any round thing or circle.
> expect, looke for.
> nationall, belonging, or consisting of a nation, or kingdome.
> sociall, ... fellowe like, one that wil keepe company, or one with whom a man may easily keepe company.
> reall, substantiall, or that is indeed subsisting.

Mulcaster (1582) had also listed *centre, national*, and *real*; Coote (1596) had included *centre, expect*, and *real*. See Schäfer (1989) for discussion and listings of words in *CTA* which antedate evidence in the first edition of the *Oxford English Dictionary*.

4 Headwords, Distribution, and Forms

Since *CTA*'s purpose was to index a selection of 'hard usuall English wordes' of French, Greek, Latin, and Hebrew origins, its overall headword profile is likely to differ from that of standard dictionaries which deal with the whole lexicon, including high-frequency core vocabulary from Germanic roots. In particular, 'hard words' might be expected to be distinctive morphologically and orthographically. The following sections try to quantify this. *CTA*'s headwords are perhaps hard *senses* rather than hard words (since Cawdrey does not explore polysemy), and this should be taken into account where my comparisons map his headwords onto whole lemmas.

4.1 Distribution across the Alphabet

Headwords in *CTA* are distributed across the alphabet in the following proportions:

A 11.5%	**F 3.2%**	**M 4.9%**	**R 5.8%**
B 2.9%	**G 2.0%**	**N 1.4%**	**S 8.3%**
C 11.9%	**H 1.8%**	**O 2.7%**	**T 3.6%**
D 8.0%	**IJ 9.7%**	**P 9.5%**	**UV 3.4%**
E 6.5%	**L 2.4%**	**Q 0.4%**	**Z <0.1%**

Table 2: Alphabetical distribution of headwords

It is always the case that different letters of the alphabet take up different amounts of a dictionary, but there is no doubt that *CTA*'s distribution is anomalous in a number of ways. Osselton (1986: 178f) points out the skewing of *A* in his discussion of *CTA* in relation to an earlier incomplete dictionary manuscript by an unknown compiler. Coote generally shows very similar skewings to *CTA*, although has an even higher percentage of words in *I*, and significantly fewer in *R* and *U/V*.

Skewings can be explored further by comparing the distribution of *CTA*'s headwords, anachronistically, with current dictionary profiles. Appendix 1 sets out two such comparisons. The first is with Thorndike's block system (see Landau 2001: 360-362, and cf. Prinsloo and de Schryver 2002: 485). This measures the likely proportions of each dictionary letter in a standard dictionary of English, produced by dividing the alphabet into 105 blocks, and allocating standardized numbers of blocks to letters; it is, however, a measure of proportions of the complete dictionary text, not just headwords. The second is with the distribution of headwords in *Collins COBUILD English Dictionary* (1985, second edition) [CCED]: another pedagogical dictionary.[3]

In comparison with Thorndike, *CTA* has an unusual number of words beginning with *A*, *I/J*, and *E*. If *CTA*'s spellings are normalized to distinguish initial *I* and *J*, then the disproportion is shown to be with *I*, which is 2.4 times greater in *CTA*. The sparsest initials represented in *CTA* are *Z, Q, G, H, B*, and, with normalized spellings, *J* and *U*. The same disproportion is evident in a comparison between *CTA* and *CCED*: more headwords in *CTA* begin with *A, E*, and *I/J* (particularly *I*), and fewer with *B, H,* and *Z*. *CTA* contains significant numbers of headwords beginning with Latinate prefixes such as *com-/con-, dis-, in-/im-, inter-, per-, pre-, pro-, re-*, and *trans-* (normalized spellings) but this does not really seem to be a factor in the uneven distribution of its headwords.

Another way of assessing the anomalousness of *CTA* might be to compare its headword distribution with that of lemmas in a corpus: what proportions of words begin with different letters, and how this varies according to frequency. Ideally, such a corpus would be of late 16th- or early 17th-century English. But in its absence, some data can be derived from a corpus of current English, and this is discussed in Section 5 in relation to Cawdrey's notion of hard words.

4.2 Lengths of Words

The headword items in this text of *CTA*, including variant spellings and forms, have a median length of 8.37 characters: 8.08 if spellings are normalized. Nearly 30% have 10 or more characters, while only around 6% have 5 or fewer. Median lengths vary according to the initial letters of headwords, and while most cluster between 7.8 and 8.5, a few lie outside this range. *B, G*, and *L* have median frequencies of respectively 7.23, 7.28, and 7.59; *C, I/J, P*, and *T* of respectively 8.98, 9.08, 8.92, and 8.81. The median length of *CTA* headwords is slightly longer than in Coote, where it is 8.2.

5 Hard Words and Related Issues

Without adequate data from late 16th- or early 17th-century English, it is difficult to assess exactly what Cawdrey's 'hard words' were, or how hard they really were for their time. What might the equivalent be in current English, and how does this relate to notions of core and non-core vocabulary? And what are today's nearest equivalents to 17th century hard words dictionaries? Since the term 'hard words' implies esoteric words and esoteric meanings, it might be tempting to suggest that these would be dictionaries of technical terms, or even dictionaries of neologisms, dialectisms, slang, or miscellaneous lexical oddities. However, in many respects, the profile of *CTA*'s lexicon seems to fit better with that of dictionaries aimed

at students, or advanced learners: that is, pedagogical dictionaries with words which are useful, perhaps have difficult meanings, but are not rare.

Cawdrey expected his users to encounter his headwords and encode them, and he excluded inkhorn terms: at the same time, they are 'hard words'. This suggests that they are less likely to be high-frequency central or core vocabulary items: in current English, this tranche of the lexicon can perhaps be quantified as the 5000 commonest lemmas.[4] An appropriate comparison, then, might be based on the next 25000 commonest lemmas in current English: compare monolingual learners' dictionaries such as *CCED* with their coverage of around 30-40000 main headword items. Estimates of adult native-speaker vocabularies vary, but figures between 30000 and 60000 are typical,[5] and so comparisons could also be made with two further frequency ranges, each of 25000 lemmas. Below this point, words occur less than once per 25 million words of (corpus) text, and are likely to be restricted to certain registers or varieties, or are simply obscure or dated: in some respects, a parallel to inkhorn terms.

Table 1 gave the frequencies in current English of *CTA*'s extant headwords. Table 3 shows how the headwords map onto four frequency bands of lemmas in the 450 million word Bank of English corpus.

CTA headwords %	*ranking*	*absolute frequencies*
22.9	**1-5000**	**6139+**
51.1	**5001-30000**	**205-6138**
12.2	**30001-55000**	**43-204**
3.7	**55001-80000**	**17-42**

Table 3: *CTA* headwords and corpus frequencies

(Of the remaining headwords in *CTA*, now with frequencies below this point, half occur just once or twice per 50 million words.) While the comparison is flawed, not least because of the gap of four hundred years, it does seem to support the idea that Cawdrey's target lexicon is not dissimilar to that of a large monolingual learner's dictionary.

These frequency bands can also be used in exploring *CTA*'s uneven alphabetical distribution of headwords, since the proportion of words beginning with particular letters varies according to frequency. Such a comparison is set out in Appendix 2.[6] Again, it is flawed because the comparison is between two very different English lexicons. Nevertheless, there is some confirmation of the skewing in *CTA*, whether overall distributions are considered, or just those of lemmas in specific bands. Curiously, *CTA*'s dense letters *A* and *E* are word initials which seem to be found less often as frequency decreases: that is, there are proportionately more words beginning with *A* and *E* in higher frequency ranges: this is also true to some extent of *I*, when spellings are normalized. The opposite is true of *CTA*'s comparatively sparse letters *B* and *G*: they are found more often as initials amongst lower frequency words. This is perhaps the reverse of what *CTA* might have been predicted to do. One hypothesis might be that the morphological patterns of common and rare words have shifted in the last 400 years; another, that Cawdrey's selected headwords – or senses – were commoner than the label 'hard words' might suggest.

6 Headwords and Word Classes

Proportions of word classes also change according to frequency: for example, rarer items are more likely to be nouns and less likely to be verbs. This is reflected in the headword lists of larger dictionaries, which normally have more nouns than smaller ones: when more words are added, they are likely to be lower-frequency nouns, such as technical terms. As a hard words dictionary, *CTA* could be expected to have more nouns than other word classes, and this proves to be the case, with approximately half of all headwords being nouns.[7] The following table shows the distribution: there are a few cases where the word class of headwords is unclear or ambiguous.[8]

nouns	**49.6 %**
adjectives	**25.5 %**
verbs	**23.9 %**
adverbs	**0.4%**
other/unknown	**0.6%**

Table 4: Word class distribution of *CTA* headwords

Word classes, however, are not distributed evenly across the alphabet. In *CTA*, proportionately more nouns start with *B* and *G*, adjectives with *F, N*, and *O*, and verbs with *D*. Proportionately fewer nouns start with *I/J*, adjectives with *G*, and verbs with *H* and *L*.

The overall distribution of word classes in *CTA* seems very similar to that for *CCED*, in spite of differences in coverage, rationale, and era. The figures in Table 5 are based on the word classes of individual senses in *CCED*, and exclude phrases and affixes:

nouns	**49%**
adjectives	**22%**
verbs	**23.9 %**
adverbs	**2.2%**
other/unknown	**2.8%**

Table 5: Word class distribution of *CTA* headwords

In comparison, corpus lemmas are distributed according to word class as follows:

	top 5K	5-30K	30-55K	55-80K
nouns	56.9%	61.7%	77.1%	86.1%
adjectives	13.9%	19%	11.4%	6.1%
verbs	19.2%	14.6%	7.6%	5%
adverbs	5.2%	4.3%	3.8%	2.7%
other	4.8%	0.5%	0.2%	0.1%

Table 6: Word class distribution of corpus lemmas

At first sight, this seems to suggest that both *CTA* and *CCED* underrepresent nouns, and overrepresent adjectives and verbs; however, polysemy needs to be factored in. Additionally, a significant number of words tagged as nouns in the corpus are names of people or places. While my calculations tried to exclude these, inevitably some remained, and these may have distorted the figures, although there is no doubt that the proportion of nouns increases as frequency falls: most rarer words are nouns.

7 Headwords and etymology

Just over 22% of *CTA*'s headwords are labelled as adopted from French (347 headwords) or from Greek (212 headwords): unlabelled items by default are from Latin. Distribution across the alphabetical text is predictably uneven. A disproportionate number of words from French begin with *A, B*, and also *L* and *R*, and disproportionately few begin with *I/J* and *O*. A disproportionate number of words from Greek begin with *E, G, P*, and in particular *H* and *M*, while disproportionately few begin with *I/J, L*, and *R*.

Nearly half of the French words are nouns, but there are more verbs and fewer adjectives than in the headword list as a whole. Most of the Greek words are nouns, and very few are verbs.

	all	*French*	*Greek*
nouns	49.6%	48.7%	80.1%
adjectives	25.5%	14.5%	16.6%
verbs	23.9%	35.9%	2.8%
adverbs	0.4%	–	–
other/unknown	0.6%	0.9%	0.5%

Table 7: Word class distribution of French and Greek headwords

While the median length of headwords in *CTA* is 8.37 characters, words labelled as Greek in origin are slightly longer (8.53 characters), and almost identical to default Latinisms (8.51 characters). Words labelled as French are shorter (7.51 characters). This in part contributes to the below-mean average length of *CTA* headwords beginning with *B* and *L*, which are disproportionately associated with French origins, and above-mean lengths of headwords beginning with *P*, which are disproportionately associated with Greek.

8 Beyond headwords

Although this paper is primarily concerned with *CTA*'s headword list, the vocabulary of its defining language must also be of interest. Words used in definitions reflect the semantic field of the kinds of items covered as headwords: common words in *CTA*'s definitions include *authority, chief, good, holy, knowledge, rule,* and *worthy*, which show up its world view, as well as general words such as *man, place, thing*, and *time*. Nearly 10% of its 'hard words' headwords also occur in definitions, some multiple times. These include topic-neutral abstract nouns such as *division, proportion*, and *value*, as well as verbs such as *confirm, consent, enlarge*, and *resist*. The most frequent definition word is *or*, largely because of *CTA*'s synonym or paraphrase method of defining:

> benignitie, gentlenes, or kindnes
> estimate, esteeme, value, or prise, thinke or iudge
> martiall, warlike, or valiant, or taking paines and delight in warres

One feature which contrasts with modern definition praxis is a lack of hedging words, such as *especially*, although a very few definitions include *et cetera*: for example, 'entra[i]ls, inward parts, as hart, liuer, &c'.[9]

9 Conclusions

While it cannot be entirely satisfactory to use modern data sources in assessing quantitatively *CTA*'s oddities, such sources do provide some confirmation of the skewing of its headword list and the selectivity of its coverage. This can be seen in the morphology of headwords, particularly what letters they begin with. It can also be seen in their grammatical classes: there are, perhaps, a surprising number of adjectives and verbs.

Any quantitative analysis of *CTA*'s vocabulary can only provide a partial description, but this can then be used as a means for more qualitative analyses. To what extent are anomalous distributions a product of the sources on which Cawdrey drew? or a product of contemporary

developments and changes in the English lexicon? Can they be related to the semantic fields of the vocabulary that he dealt with, and does this help understand the world view that he was projecting? And what, ultimately, were its hard words? Revisiting Cawdrey's *A Table Alphabeticall*, four hundred years after it first appeared, shows up how much has changed in lexicographical method – and equally how issues of headword selection and consistency remain.

Endnotes

1 See also work by Siemens (1994 and elsewhere) and his computer-assisted explorations of the methodology and composition of *CTA*, including a comparison of its different editions.
2 Corpus data is drawn from the Bank of English corpus created by COBUILD at the University of Birmingham.
3 COBUILD data is cited by kind permission of HarperCollins Publishers.
4 Other estimates of core vocabulary might be lower than this: see Nation and Waring (1997) for discussion in relation to second language learning.
5 See Aitchison (1987: 5ff) for discussion of vocabulary size in relation to native speakers.
6 Prinsloo and de Schryver (2002) report on a similar exercise, using data from the British National Corpus, in the context of developing a methodology to ensure evenness of coverage across the alphabet, with particular reference to English and Afrikaans dictionaries. There are slight discrepancies between their corpus-derived figures and those set out here in Appendix 2: this is likely to be because of differences in size and composition of the two English corpora, and different lemmatization principles.
7 *CTA* itself contains no grammatical information: this data is based on word forms and definitions.
8 Siemens (1994) has different statistics for adjectives and verbs, because of different criteria for grammatical classification.
9 See Siemens (1994) for further discussion of the structures of definitions.

References

A: Dictionaries

Cawdrey, Robert. 1604. *A Table Alphabeticall.* Gainsville, Fla: Scholars' Facsimiles and Reprints.
Coote, Edmund. 1596. *The English Schoole-maister.* Menston: Scolar Press.
Collins COBUILD English Dictionary. 1995, ed. 2. London and Glasgow: HarperCollins.
Mulcaster, Richard. 1582. *The First Part of the Elementarie.* Menston: Scolar Press.
Oxford English Dictionary. 1884-1928, ed. 1. Oxford: Oxford University Press
Thomas, Thomas. 1587. *Dictionarium Linguae Latinae et Anglicanae.* Menston: Scolar Press.

B: Other works

Aitchison, J. 1987. *Words in the Mind.* Oxford: Blackwell.
Green, J. 1996. *Chasing the Sun: Dictionary-makers and the Dictionaries they Made.* London: Jonathan Cape.
Hayashi, T. 1978. *The Theory of English Lexicography 1530-1791.* Amsterdam: John Benjamins.
Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.
Nation, P. and Waring, R. 1997. 'Vocabulary Size, Text Coverage and Word Lists' in N. Schmitt and M. McCarthy (eds.) *Vocabulary: Description, Acquisition, and Pedagogy.* Cambridge: Cambridge University Press. pp 6-19.
Osselton, N. 1989. 'Alphabetisation in Monolingual English Dictionaries to Johnson' in G. James (ed.), *Lexicographers and their Works.* Exeter: University of Exeter Press. pp 165-173.

Osselton, N. 1986. 'The First English Dictionary? a Sixteenth-Century Compiler at Work' in R. Hartmann (ed.) *The History of Lexicography: Papers from the Dictionary Research Centre Seminar at Exeter, March 1986*. Amsterdam and Philadelphia: John Benjamins. pp 175-184.

Prinsloo, D., and de Schryver, G.-M. 2002. 'Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English' in A. Braasch and C. Povlsen *Proceedings of the Tenth EURALEX Congress*. Copenhagen: Center for Sprogteknologi. pp 483-494.

Schäfer, J. 1989. *Early Modern English Lexicography*. Oxford: Oxford University Press.

Siemens, R.G. 1994. 'The Acorn of the Oak: a Stylistic Approach to Lexicographical Method in Cawdrey's *A Table Alphabeticall*'. http://www.chass.utoronto.ca/epc/chwp/siemens1, 23/07/03.

Starnes, De W.T., and Noyes, G.E. 1991. *The English Dictionary from Cawdrey to Johnson*. Amsterdam and Philadelphia, John Benjamins. (originally published 1946, Chapel Hill: University of North Carolina)

Stein, G. 1985. *The English Dictionary before Cawdrey*. Tübingen: Niemeyer.

Stein, G. 1986. 'Sixteenth-Century English-Vernacular Dictionaries' in R. Hartmann (ed.) *The History of Lexicography: Papers from the Dictionary Research Centre Seminar at Exeter, March 1986*. Amsterdam and Philadelphia: John Benjamins. pp 219-228.

Appendix 1

The following table compares *CTA*'s headword distribution with the proportion allocated by Thorndike to each letter of the alphabet, or set of letters, and with the headword distribution in *CCED*. The columns headed 'difference' give very approximate measures of any differences or disproportions in *CTA*, computed by dividing *CTA*'s figure with Thorndike's or *CCED*'s. For example, *A* in *CTA* is twice the Thorndike allocation and twice *CCED*.

letter	*% CTA headwords*	*% Thorndike blocks*	*difference*	*% CCED headwords*	*difference*
A	11.5	5.7	2	5.8	2
B	2.9	5.7	0.5	5.9	0.5
C	11.9	9.5	1.3	10.0	1.2
D	8.0	5.7	1.4	6.0	1.3
E	6.5	3.8	1.7	3.8	1.7
F	3.2	4.8	0.7	4.6	0.7
G	2.0	3.8	0.5	3.2	0.6
H	1.8	3.8	0.5	3.8	0.5
IJ	9.7	4.8	2	5.0	1.9
K	–	1.0	n/a	0.6	n/a
L	2.4	3.8	0.6	3.2	0.8
M	4.9	4.8	1	4.8	1
N	1.4	1.9	0.7	2.0	0.7
O	2.7	2.9	1	2.6	1.1
P	9.5	7.6	1.2	8.5	1.1
Q	0.4	1.0	0.4	0.4	0.9
R	5.8	4.8	1.2	5.4	1.1
S	8.3	12.4	0.7	12.1	0.7
T	3.6	4.8	0.8	4.9	0.7
UV	3.4	3.8	0.9	4.2	0.8
W	–	2.9	n/a	2.8	n/a
XYZ	<0.1	1.0	<0.1	0.5	<0.1

Appendix 2

This table sets out percentages of lemmas in the Bank of English within four frequency bands, according to their initial letters: lemmas are distinguished according to word class, so nominal and verbal forms are counted separately. The average percentage for the top 80000 lemmas is also given. Percentages for headwords in *CTA* are given in the last two columns.

initial letter	% top 5K	% 5-30K	% 30-55K	% 55-80K	% average	CTA letter	% headwords
A	6.8	6.1	5.6	6.1	6.0	A	11.5
B	5.2	6.1	6.4	6.7	6.3	B	2.9
C	9.9	9.2	8.8	8.1	8.8	C	11.9
D	5.6	5.9	5.0	5.0	5.3	D	8.0
E	5.2	3.8	3.6	3.6	3.8	E	6.5
F	4.8	4.5	4.0	3.8	4.1	F	3.2
G	2.6	3.5	4.0	3.9	3.7	G	2.0
H	3.0	4.0	3.9	4.2	4.0	H	1.8
I	3.8	4.1	3.2	3.1	3.5	IJ	9.7
J	1.0	1.0	1.0	1.3	1.1		
K	0.7	1.0	2.0	2.9	1.9	K	
L	3.7	3.2	3.6	3.6	3.5	L	2.4
M	4.7	5.8	6.5	6.8	6.2	M	4.9
N	2.3	2.2	2.7	3.0	2.6	N	1.4
O	2.8	2.4	2.5	2.4	2.4	O	2.7
P	7.8	7.3	7.4	6.7	7.2	P	9.5
Q	0.4	0.5	0.7	0.6	0.6	Q	0.4
R	6.2	5.4	4.6	4.3	4.9	R	5.8
S	11.3	11.1	10.6	10.1	10.7	S	8.3
T	5.8	5.0	5.3	5.4	5.3	T	3.6
U	1.3	3.0	2.9	2.6	2.7	UV	3.4
V	1.3	1.7	1.5	1.8	1.6		
W	3.0	2.7	3.0	2.7	2.8		
X	<0.1	<0.1	0.1	0.1	0.1		
Y	0.6	0.4	0.5	0.6	0.5		
Z	0.1	0.3	0.5	0.6	0.4	Z	<0.1