WORD FREQUENCY AND KEYWORD EXTRACTION

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

Word Frequency: Use or Misuse?

John M. Kirk, *Queen's University Belfast, Northern Ireland.*

Keywords
word frequency, corpora, numbers , contextualization, replicability , orthographic, phonological, morphological, lexical, grammatical, onomastic , lexicographical. statistical, numeral, discourse,

Abstract
This paper will not be concerned with statistical treatments of word frequency beyond percentage distributions and relativized frequencies per thousands or millions of words. Its primary concern will be with frequency as a property of data, adopting a critical look at statements such as 'each text comprises 2,000 words'. It will be concerned with words as tokens, types and lemmatised types; the range of functions and meanings of words; and words and lexemes. It will consider words of low frequency as well as of high frequency.

In its critical section, it will ask whether word frequencies are self-explanatory or need explanation, and whether approximation is as useful as precision. It refers to a range of well-known corpora of English as well as the three corpora which I have compiled: *Corpus of Dramatic Texts in Scots*, the *Northern Ireland Transcribed Corpus of Speech (NITCS)*, and the Irish component of the *International Corpus of English (ICE-Ireland)*.

Summary

- Word frequency is the placing of numbers on language or the representation of language through numbers
- Word frequency provides an instantiation of the claim that 'linguistics is the *scientific* study of language'
- Word frequency promises precision and objectivity whereas the outcome tends to be imprecision and relativity
- Word frequency is not an end in itself but needs interpretation through contextualisation whence the relativity and comparison
- Word frequency is not science but methodology which lends itself to replicability.

One of the aims of this paper is to deconstruct statements of the following type: 'each text contains (approx.) 2,000 words', in which there are two issues: the concept (word) and the number (2,000).

Classes of Words

Of the many subclassifications of words, one which might suit our present purposes is the taxonomy proposed by Tom McArthur (1992)[1] which offers eight possible word classes:

1. The orthographic word
2. The phonological word
3. The morphological word
4. The lexical word
5. The grammatical word
6. The onomastic word
7. The lexicographical word
8. The statistical word

To this list, I wish to add a further two classes:

9. The numeral word
10. The discourse word

Of those eight or ten types, it is class 8. – the statistical word – which is usually associated with the notion of word frequency. McArthur provides the following definition:

> " … word in terms of occurrences in texts is embodied in such instructions as 'count all the words on the page': that is, *count each letter or group of letters preceded and followed by a white space*. This instruction may or may not include numbers, codes, names, and abbreviations, all of which are not necessarily part of the everyday conception of 'word'. Whatever routine is followed, *the counter deals in tokens or instances* and as the count is being made the emerging list *turns tokens into types*: or example, there could be 42 tokens of the type *the* on a page, and four tokens of the type *dog*. Both the tokens and the types however are *unreflectingly* spoken of as words." (*OCEL* 1992; reprinted McArthur 1999: 47) (my emphases)

Statistical words are words or any string of characters bounded by space which can be counted by a computer. No other distinction is made. Such words are regarded as word 'types'.

When the statistical word test is applied to ICE-Ireland[2], what frequency precision do we find? For the present, all figures are based on the beta version of the spoken component.
It is regularly stated that the spoken component of an ICE corpus comprises 300 texts each of 2,000 words, thus amounting to 600,000 words in total. In the case of ICE-Ireland, the total is 623,351 words comprising 300 texts ranging from 960 to 2,840 words each. Whereas these totals already exclude mark-up, they still include X-corpus, editorial comments and partial words (marked up as <.> … </.> and underlined here for presentation), as shown in 1. and 2.:

1. Uhm Marie-Louise and I were in you know the Bang <,> and <{> <[> <u><.> Oluf </.></u> <#> What is it <#> Olufsen </[>
2. And uh <,> like three thousand eight hundred <#> And there was another one at four <u><.> hu </.></u> four thousand two hundred and something

The question thus arises whether, in terms of McArthur's taxonomy, those 623,351 statistical words are also 623,351 orthographic words, or 623,351 phonological words, or even 623,351 morphological words. They are not 623,51 lexical words (in the sense of lexical types), even less 623,351 lexemes (where *die*, *pass on* and *kick the bucket* are considered single lexemes).

Let us consider briefly each type of word in turn.

1. The orthographic word
One instance of an orthographic word is where the word has dual spellings, as in:

> *airplane, aeroplane; esthetic, aesthetic; archeology, archaeology; connection, connexion; counselor, counsellor; gray, grey; instill, instil; jeweler, jeweller; jewelry, jewellery; libelous,*

libellous; marvelous, marvellous; mollusk, mollusc; mustache, moustache; panelist, panellist; paralyze, paralyse; analyze, analyse; pajamas, pyjamas; skeptic, sceptic; color, colour; honor, honour; labor, labour; traveler, traveller; traveling, travelling; willful, wilful; woolen, woollen.

These are well-known standardized instances of dual spellings which as a result of the institutionalisation are regarded as 'American' or 'British'. When we investigated those spellings in the written texts of ICE-Ireland, all but a few which had been published in Ireland, we found that ICE-Ireland is actually more British than ICE-GB, as shown in Table 1!

Table 1: Verbal Spellings in *–ise* and *–ize*

	ICE-NI	ICE-ROI	ICE-GB	Total
-ise	17	9	35	61
-ize	2	1	12	15

Dialect words present another particular instance of the orthographic word as many such words have survived in oral currency and have never had a standardized form. In Ireland, there are many words for the national crop, the humble potato, which can be listed under the headword *potato*, as in the *Concise Ulster Dictionary*:

***Potato*:** the national crop in all parts of Ireland: *potato, pitatie, pirtie, pirta, purta, purty, pitter, porie, pratie, praitie, prae, prata, prater, pritta, pritty, pruta, poota, tater, tattie, totie* . (Hiberno-English forms are recoded as *pratie, praitie*, etc.; Scots forms as *pitatie, tattie, tottie*; and a southern English form as *tater*).

For other words, there is no agreed standardised form, as in the various forms of the dialect word for 'embers' borrowed from the Irish word *griosach*:

greeshoch, greesagh, greesach, greesay, greeshagh, greeshaugh, greeshaw, greesha, greshia, greeshy, greesh, grushaw (< Irish Gaelic *griosach*) cf. *greeshog, greesog, greeshock* (Irish Gaelic *griosog*, 'small flying embers').

Some words are harder to identify. The word for 'twilight' or 'dusk' is *dailygan* in the Scots dictionary of Ulster, but the *Concise Ulster Dictionary* lists:

daylight going, daylit goin, dayligoin, daylight gone, dayligone, dailagone, dailygan, dayligane, dayagone, dayligo

making it unclear where the underlying base form is 'daylight going' or 'daylight gone'.

As statistical words, these orthographical words would be counted separately – as types – whereas they merely represent various pronunciation variants of the same lexical type. Each of these three lists present only one lexical type.

The BBC is currently running a nationwide dialect project called *Voices*. It falls into this same trap of counting orthographic variants as separate – it goes so far as to say *unique* – words. The Voices website (www.bbc.co.uk/voices/) states:

"The Word Map has been highly successful; an initial look at the data suggests 32,000 users have registered […] *23,000-odd unique words (including spelling variations)* …"

An instance of this practice is shown in the Appendix/handout – the 'words' elicited by the investigation into words for 'attractive'.

With regard to the issue of word frequency, orthographic words present as many difficulties as statistical words.

2. The phonological word

Phonological words are conceivable as several subtypes: vocalised words (*phoar* words in the appendix), partial words (the initial segments of a word but not the complete word, as in 1. and 2. above), orthographic or prounounciation-variable words (presenting different pronunciation variables, as in *economics* or *tomato* or because of shifting stress positions in *controversy* or in the dialect forms above), syllabic words (e.g. *gonna, hadda, musta, needti, wanna*, etc.) or even clausal or intonation-unit words (e.g. *spindona*, *fellafellafella, gerritupyi*, etc.) In these ways, phonological words either become orthographic words (which in turn become statistical words) or appear as conflated words which, if counted as statistical words, under-represent the actual total. There is no corpus of segments or syllables, although, interestingly, it is claimed by David Crystal (2003) that 25% of speech is made up of only 12 syllables.

So with regard to the issue of word frequency, phonological words present many other difficulties too.

3. The morphological word

Morphological words may be lexical or grammatical words. First, let us consider lexical morphemes. The prefixes: *cyber-*, *e-*, *eco-*, *euro-*, etc. all became frequent … as a result of change in technology, politics or attitudes to the environment. The sudden increase of use of such forms helps to construct the discourses about these new realities. As the *Oxford Dictionary of Ologies and Isms* shows, many prefixes and suffixes are specific to particular domains – even linguistics can claim *glosso-, grapho-, logo-, semio- Slavo-* as prefixes and *–eme, -gram, -graphy, -lect, -lepsis, -logue, onym, -phasia, -speak and –word* as suffixes.

In the southern component of ICE-Ireland, we discovered that clipped words with the suffix *–o* marked colloquial speech, perhaps even slang:

Defos	S	Slang; 'definites' (replies, etc.); (X OED, SUE)
Invos	S	Slang 'invitations'
Morto	S	Slang 'mortified'
Séamo	S	Form of *Séamus*
Smarmo	S	In ICE as interjection (< *smarmy*)

Other forms were:

Relies	S	Slang 'relatives'
Sca	S	Slang 'news, gossip' (< *scandal*); (X SUE, OED)

There are no such forms in ICE-GB[3]. Even if the absolute numbers are few, their presence in one corpus and not in another may be interpreted as significant – indicative of innovating collquialisms, possibly slang words. The more lexical items adopting this *–o* suffix, the more the pattern becomes established. Frequency can thus reveal cultural innovation.

Grammatical morphemes offer numerous challenges. Some are mere variants of a single form, sometimes conditioned by external factors, such dialect contact in the case of the past tense form of *bring* as *brought* or *brung*, or a negated form of *could* as *couldn't* and *couldnae*. ICE-Ireland has six instances of *gotten* alongside *got*, each with a clear dynamic meaning. Grammatical variants and grammatical innovations may be interpreted in terms of external contexts, but they may also be indicative of changes in the particular sub-system itself. The form *gonna* may be as the output of a grammaticalised progressive *go* construction, but only if *gonna* is transcribed as such – in ICE-Ireland, it was not so transcribed – only the standard *going* was used for every instance of progressive *go* – in stark contrast to the British National Corpus where its inclusion was left to the subjective preference of the audio-typists who were transcribing the tapes.

When the statistical word test is applied to grammatical words, the result can be confusing. *is* and *was* are often shown to rank among the most frequent words, but they are only verb forms - they are neither verb types nor the most frequent verbs (for that we need the total of all forms of *be*). Although Table 2 presents frequencies of individual forms of *be* in three spoken corpora, which show some consistency across corpora for each form, it does not show the frequency of *be* itself.

Table 2: Frequencies of BE Forms

Form	ICE-IRL Spoken	LONDON-LUND	MILLER
	f.	f.	f.
-'s	17.21	21.64	30.60
Is	9.46	10.45	7.05
was	10.64	10.52	7.51
be	6.15	5.46	3.44

Frequencies of *will* require the sum of *'ll, will, won't* and whatever other spelling variants are used. So, when it comes to word frequency, a lot of caution and qualification is needed around the frequency of grammatical words.

4. The lexical word

As already shown, lexical words are often mistaken for variants of their realization: phonological words, which are rendered in writing as orthographic words, or morphological words (particularly with different noun or verb forms). Much of the interest shown in lexical words is as lexical types, or lemmatised types, not as families of realizations. Even if we establish frequencies for lexical types – something the statistical word does not do – how are we to interpret the result? As Shakespeare wrote, 'What's aught but as 'tis valued?*'* Of the many possible contexts, I raise only three here: semantic prosody, attitude raising, and constructions of identity or reality.

Semantic prosody

Following the pioneering work of Bill Louw (1993),[4] the notion of semantic prosody is now generally accepted. *Utterly* is regarded as having a negative prosody, i.e. it collocates with words expressing a negative meaning, so that, in ICE-Ireland, we find that prosody confirmed in the six examples: *utterly boring, utterly unacceptable; condemn utterly (x2); utterly/serial killers* (ICE-Ireland).

Attitude raising

The common word *happy* seems innocent enough until put into the literature for boy scouts and girl guides by Lord Baden-Powell, who urges that the purpose for girls in life is to make boys happy, whereas the purpose of boys in life is simply to be happy. The accumulation of overuse in those texts is shown by Stubbs (nn) to turn the word *happy* into a sexist term.

Constructions of identity or reality

In a masterly study of keywords, Paul Baker (2004) shows how gay identity is constructed very differently by different groups of people. For the House of Lords, key words for the pro-formers were *law, rights, sexuality, reform, tolerance, orientation, sexual. human***,** whereas key words for the anti-reformers were *buggery, anal, indecency, act, blood, intercourse, condom***.** For the British tabloid press covering crimes on gay men, key words were *transiency, acts, crime, violence, secrecy, shame, shamelessness, promiscuity***.** In contact ads, British gay men described themselves as *guy, bloke, slim, attractive, professional, young, tall, non-scene, good-looking, active, caring, sincere***.** In gay fantasy literature, gay men are described as brutes **(***socks, sweat, beer, football, towel, team***)** or emotionless machines **(***lubed, jacked, leaking, throb, throbbing, spurt, spurts, pumped, pumping***).** In safer sex awareness leaflets distributed to gay men, gay men are described as animals, and gay sex as violent **(***grunted, groaned, grabbed, shoved, jerk, jerked, jerking, slapping, pain***);** at the same time, gay men's language is shown to be informal, non-standard and impolite **(***fucker, cocksucker, faggot/fag, stuff, yeah, shit, hell, fuckin, ain't, wanna, gotta, gonna, 'em, kinda, real, hey, damn, good***).** In each of these settings, it was the frequency of words occurring above the norm that created the very different discourses in each context and foregrounded the perspective or point of view.

Baker shows convincingly and beyond doubt that by studying word frequencies common, everyday words words like *human* or *young* when over-used – and thus with a relatively high frequency – in particular texts become keywords and agents in the creation of and also discrimination between those discourses.

Similarly, ICE-Ireland creates Ireland – through the use and frequency of various classes of lexical words: dialect words, Irish loanwords, other words in ICE-IRL deemed 'Irish' and institutional words (many of them onomastic words) – that frequency always being relatively high compared to other ICE-corpora, where such items do or will not occur. These words are keywords because neither is used in any other ICE corpus nor could they construct anything but Ireland. An analysis of a sample of ICE-Ireland spoken component revealed the following Irishisms, as listed in Tables 3 and 4, where 'N' and 'S' relate to northern and southern distributions:

Table 3: Irish loanwords

Fleadh	N, S	Traditional music festival (< Irish)
Gaeltacht	S	Irish-speaking district (< Irish)
Poitín	S	Illicit distilled spirits (< Irish); cf. OED
Scór	S	'Tally' (<Irish)

Table 4: Lexicon of other words in ICE-IRL deemed 'Irish'

Maracycle	N	Long distance cycling race
Motorsports	N	Sports using cars or motorcycles (X OED)
Bogger	S	HibE dialect 'person from rural areas' (cf.OED)
Feck	S	Slang; variant of *fuck* (X SUE)
Greenkeeper	S	One who maintains a green (X OED)
Imprimatured	S	Use of imprimatur as verb; X OED
Knacker	S	Derogatory for 'traveller'; extended to more general derogatory sense; cf. indirectly related senses in OED, SUE
Legger	S	did a legger 'ran away'; X SUE, cf. OED Sc. and dial. *leg* 'use the legs, to walk fast or run'.
Liveweight	S	Weight of live animal (X OED)

A knowledge of these rather unexceptional words – English in form, but only used in Ireland – are important for a knowledge of Ireland.

5. The grammatical word

Grammatical words are not morphological words which are grammatical variants of a lexical word. Although sometimes homonyms, grammatical words are grammatical in their own right, although they may be realised by a subset of phonological/orthographic variants, as in the case *'ll* for *will* or *won't* as a contracted form of negated *will*. Although benchmark frequencies are sometimes given – *The Comprehensive Grammar of the English Language* asserts that *will* occurs four times in every 1000 words, I find that in Scottish English (Scots) it occurs at least eight times in every 1,000 words, the figure attributable not simply to additional functions but through the restructuring of exponents within the system of modality – in effect because of internal systemic differences. However, in the Northern Ireland Transcribed Corpus of Speech (NITCS),[5] the most frequent modal is *would*, for which the explanation is external. Although *would* carries a habitual-in-the-past meaning in standard English, Irish English marks habituality in the present, under the transfer of that category from Irish through language contact. Moreover, the majority of the texts in NITCS constitute interviews about childhood reminiscences and reflections on changes experienced by the interviewees during their lifetime. So, contextually, it is not surprising that *would* appears so frequently. These results are presented in Table 5 (based on Kirk 1986).[6]

Table 5: Frequency of Modal Verbs

	NITCS	ICE-IRL		LONDON-LUND	MILLER[7]
	F	N	F	F	F
WILL	1.77	2584	4.14	4.28	6.19
WOULD	9.69	3652	5.85	3.51	4.79

Grammatical words not only express a range of meanings, they can occur in a range of syntactic constructions, which may in some cases correlate with particular meanings. High

frequencies of *get* as a lexical word (as in Table 6) hide the many possibilities both for complements and for premodification through auxiliary, modal or catenative verbs.

Table 6: Occurrences of get

	N.	F/1000
NITCS	2397	10.06
ICE-IRL Spoken	3005	4.92
GLASGOW[8]	723	9.65

By contrast low frequencies of the *after*-perfect construction can be shown to correlate directly with different contexts:

After-perfect ICE-IRL: all 9 are *southern*
After-perfect in NITCS: majority (3/5) among *Catholic* speakers
After-perfect in GLASGOW: all 9 are ethno-linguistic markers - by the same *Catholic* speaker

The form-function relationship which lies at the core of grammatical words further complicates the issue of word frequency.

6. The onomastic word

As already indicated, institutional names in ICE-IRL act as key words in the construction of Ireland. But onomastic words raise issues with regard to statistic words … how many words are in a single name? How many ways are there to spell particularly an Irish name? Onomastic names may also occur as acronyms. Here is a list of onomastic words from ICE-Ireland ('N' and 'S' again denoting northern and southern distributions in ICE-NI and ICE-ROI respectively).

Table 7: Local names (or onomastic lexicon)

Aer Lingus	N, S	Irish national airlines (< Irish)
Radio Telefis Éireann	N, S	RTÉ; the Irish broadcasting authority
Gardaí	N, S	Plural of *garda*, member of Garda Siochána (<Irish)
Taoiseach	N, S	Prime minister in Irish government (< Irish)
DENI	N	Department of Education Northern Ireland
DHSS	N	Department of Health and Social Services
UUP	N	Ulster Unionist Party
UYO	N	Ulster Youth Orchestra
Forum	N	Northern Ireland Forum for Peace and Reconciliation
RUC	N	Royal Ulster Constabulary
EHSSB	N	Eastern Health and Social Services Board
An Bord Pleanála	S	The Irish planning authority (< Irish)
Ceann Comairle	S	Presiding officer of the Dáil (< Irish)
Coláiste Íde	S	[< Irish; name of local school]
CRC	S	Central Remedial Clinic
Cultúrlann na hÉireann	S	[< Irish; name for Irish traditional culture centre]
Dáil	S	Dáil Éireann; the main Irish legislative body (<Irish)
EIS	S	Environmental Impact Statement
Fáinne	S	Lapel pin associated with speaking Irish (<Irish 'ring')
Telecom Éireann	S	Irish national telephone company (< Irish)
Fás	S	Irish national employment agency (<Irish)
Fianna Fáil	S	Irish political party (<Irish)
Féile	S	In ICE refers to specific music festival (< Irish)
Garda Siochána	S	Irish police force (<Irish)
Oireachtas	S	Combined houses of the Irish parliament (< Irish)
PRSI	S	Pay-Related Social Insurance
RTC	S	Regional Technical College

Seanad	S	The upper house of the Irish legislature (< Irish)
Tánaiste	S	Deputy prime minister in Irish government (< Irish)
Toisigh	S	Plural of Taoiseach
TD	S	Member of Dáil (< Irish *Teachta Dála*)

7. The Lexicographical word

The lexicographical word adopts a different approach to word frequency. Dictionaries attempt only to reflect reality of use, so that some frequency information is provided implicitly through the display of spelling variants (as shown above) or different senses (the more senses a word has, the more frequent it is likely to be – cf. *peregrinate*, which has one basic sense, with *to go on*, with 14 different senses). Although nowadays, lexicography is heavily corpus-based, the inclusion of frequency information in a dictionary implies a certain predictability about what the dictionary user is likely to find.

Lexicographical words are headwords. Most are lexical words; some are grammatical words. They are not orthographic words, nor morphological words (although there is a debate in EFL circles about the choice of headword for verbs in early learner dictionaries given that past tense forms can be more frequent than base forms (e.g. *declined*, that well-known example discussed by John Sinclair). Some headwords proliferate numbered subdivisions – either on the basis of word class e.g. *round* has several numbered entries (as if separate words are created through polysemy) as in *Macmillan* but is listed as a homograph (i.e. one headword, with several subsenses) as in *Collins English Dictionary* or *Encarta World English Dictionary*. Many headword lists also include onomastic words. Thus choice and nature and therefore frequency implications of headwords have long been controversial.

There is also separate frequency issues with regard to lexicographical words. Current dictionaries have established frequency tables with regard to content, similar to the literature on corpora size. Consider Table 8:

Table 8: Dictionaries and Word Frequency

Dictionary	Headwords	References	Text
Encarta	100,000	400,000	3.5m
NODE		350,000	4m
Collins 4E		180,000	3.6m

All the same, it is hard to know what exactly each dictionary is counting. The *Concise Ulster Dictionary* boasts 'over 15,000 words' on its cover, but in fact there are 19,936 headword entries.

The issue of word frequency raises the question of a frequency dictionary, of which it could be claimed there already are several, especially:

- K. Hofland and Stig Johansson, *Word Frequencies in British and American English* (The Norwegian Computing Centre for the Humanities, Bergen, 1982)
- W. Nelson Francis and Henry Kucera, *Frequency Analysis of English Usage Lexicon and Grammar* (Houghton Mifflin, 1982)
- Knut Johansson and Knut Hofland, *Frequency Analysis of English Vocabulary and Grammar* Vol. 1: Tag Frequencies and Word Frequencies, Vol. 2: Tag Combinations and Word Combinations (OUP, 1989),
- Gregory James, et al. *English in Computer Science: A Corpus-based Lexical Analysis* (Longman Asia, 1994)

Frequency information, however, is already accommodated in several dictionaries using different methods:

- The number system (frequency is indicated by numbers, as shown in frequency dictionaries)
- The star system (frequency is indicated by stars)
    - *COBUILD Dictionary of Idioms:* *** = 1/2m; ** = ?; * 1-3/10m

    - *Macmillan English Dictionary: For Advanced Learners*: * 'fairly common'; *** 'one of the most basic words'
- The colour system (frequency is indicated by different headword colours)
    - *Longman Dictionary of Contemporary English, 4th edn.* (red), *Macmillan* (red)
- The graph system (frequency is indicated by distributional graphs)
    - *now* (speech vs writing) in *Longman Dictionary of Contemporary English, 3rd edn*
    - *notice* (types of complement) in *Longman Dictionary of Contemporary English, 4th edn*

The result is that, on the basis of corpus research, and by treating the item as a closed system within which the variation occurs, both distributional and benchmark frequency information is given.

Lexicographers remain cautious about using word frequency as the basis of headword choice, placing their doubt on the adequacy of any sample or choice of corpus material to be a reliable indicator of frequency, and on the methodology to be sophisticated enough to reflect not only form frequency but in the case of polysemy sense frequency too. According to the American doyen of lexicography, Sidney Landau (1984),[9] the *The American Heritage Word Frequency Book* based on the Brown corpus and compiled by Francis and Kucera (mentioned above) is 'valuable but flawed' because a corpus of 1,000,000 words is "far too slight to give any true indication of the frequency relationships of the entire lexicon. … A statistically-useful corpus would have to be many times larger than 5 million words."

Nevertheless, according to Landau:

> "[there is] a sense in which dictionaries do use frequency counts – that of their own citation files. A dictionary citation file is a collection of quotations of actual usage selected to serve as a basis for constructing definitions or for providing other semantic or formal information (such ad collocation, degree of formality, spelling, compounding, etymology, or grammatical data). Citation files may also include transcriptions or recordings of spoken forms. The manner of collection and use of citation files for defining will be discussed in the next chapter. Suffice it to say here that as traditionally collected, citation files, however vast, and Merriam-Webster's files reputedly number over 12 million – have been assembled in too haphazard a manner to be used as a reliable guide to frequency. As James A.H. Murray had occasion to remark in connection with the OED files, citation readers all too often ignore common usages and give disproportionate attention to uncommon ones, as the seasoned birdwatcher thrills at the glimpse in the distance of a rare bird while the grass about him teems with ordinary domestic varieties that escape his notice." (Landau 1984: 79-80)

9. The numeral word

To McArthur's taxonomy, I wish to add two new classes of words. As a transcriber of speech and compiler of corpora, I'm aware that these two classes reveal themselves in high numbers and present their own difficulties with regard to form and thereby word frequency.

Numbers and numeral words are neither lexical words nor grammatical words. They fall between classes. There are always difficulties of transcription as many utterances are no more than a spoken version of a number which exists primarily in writing. It is interesting that some of the most recent dictionaries include a section on 'Numbers that are used as words' (*Cambridge Advanced Learners' Dictionary*) and 'Numbers that are entries' (*Macmillan English Dictionary: For Advanced Learners*). But even in grammars, the treatment of numbers and numerals can vary, as between inclusion in a chapter on 'giving information about people and things' (*The COBUILD English Grammar*) and a chapter on 'lexical word-formation' (*The Cambridge Grammar of the English Language*).

10. The discourse word

In discourse, words can have a function which is neither lexical nor grammatical, which may also be pragmatically indeterminate, but which nevertheless is relevant for the development and cohesion of the conversation. In the prosodically and pragmatically tagged version of ICE-Ireland,[10] we have identified such discourse words as 'Indeterminate Conversationally-relevant Utterances' or <icu>, as in the following example:

(3) <$A> <#> <exp> I 'm not even sure 2exActly when I 'll 2nEEd somebody from% </exp>
<$B> <#> <icu> 2Right% </icu>
<$A> <#> <rep> But uhm I would need an 1Extra pair of 2hAnds% </rep>

The category includes backchannels. Although discourse words will not be distinguished as statistical words from any other type, a considerable proportion of any spoken text comprises <icu> utterances.

Discourse markers are also marked in ICE-Ireland. Three types are identified: syntactic, lexical and phonological. Multi-word discourse markers are hyphenated to distinguish their functional use; all discourse markers are asterisked.

Table 9: List of Discourse Markers in the PPD Corpus

Syntactic	Lexical	Phonological
Do-you-know		
Do-you-see	Ah-no	Ah
I'd-say	Ah-well	Arrah
I-know	Ah-right	Och
I-mean	Actually	Oh
I-see	All-right, alright	
I-suppose	God, Jesus, Jeez	
I-think(-that)	Just	
You-know	Kind-of, kinda	
You-see	Like [focus]	
See	My-God, My-gosh	
	No, naw, no-no	
	Now	
	Oh-God, Oh-gosh	
	Oh-my-God, God-Almighty	
	Oh-Jesus	
	Oh-right	
	Oh-well	
	Oh-yeah, Oh-yes	
	Okay	
	Only	
	Right	
	So	
	Sort-of, sorta	
	Sure	
	Then	
	There	
	Well	
	Yeah-no	
	Yeah-yeah	
	Yes, yeah, yup, aye	

Transcribers of speech need to think through their policy for encoding such speech features. Whatever decision is taken, there are implications for word frequency. The London-Lund corpus and the Corpus of American Spoken English transcribes non-lexicalised, non-grammaticalised sounds phonetically, e.g. with the phonetic symbol schwa. Regardless of vowel quality, in ICE-Ireland such vocalized sounds used as hesitation markers or fillers are transcribed uniformly as *uh* and *uhm* (depending on whether there was a final audible nasal).

Ten classes of word frequency?

These ten classes of words offer themselves as ten classes of word frequency. What the examples have shown is that, in each word type, frequency is a clear factor. Frequency becomes a factor when a link is inferred between the frequency and the context. The context may be the linguistic system itself (as with exponents of modality or cases of grammaticalisation); and the context may be external, which can be interpreted as conditioning the frequency and so the pattern of frequency variation of which it is a part. External factors are many and varied – they may have to do with speakers (whether identified by country, province, region, age, sex, sexual orientation, education, life-history, L1/L2 speaker, etc.), or discourse situations (whether what is spoken is read, prepared, spontaneous, broadcast, or what the audience or purpose or intended effect might be). It may have to do with the method of recording (ICE is fly-on-the-wall without fieldwork presence; NITCS is wholly driven by fieldworker questions), the time of recording as a special moment in history; or the discourse which comes to be constructed, intentionally or otherwise.

Comparing frequencies in corpora

Comparing corpora will always generate different frequencies for interpretation by such external conditioning factors:

- ICE-Ireland vs. ICE-GB vs. ICE-(whatever)[11]
- ICE-NI vs. ICE-ROI
- ICE-NI Spoken vs. NITCS
- ICE(-whatever) vs. LLC[12] vs LOB[13] vs. FLOB[14], etc.
- ICE(-whatever) vs. BNC
- CDTG (GLASGOW)[15] vs. Leuven[16]

Comparing corpora also generates the need to go beyond raw frequencies – numbers of occurrences – and relativise frequencies (per 1,000 or 10,000 or 100,000 or even 1,000,000 words to compare occurrences from corpora or datasets of different lengths) as one of a closed set (percentage distribution), sometimes to stand as a benchmark figure relativised to 1,000 words or 1million words.

Comparing corpora finally depends on replicability. The statistical word test may seem the obvious and easy answer. But, as the present examples show, the significance of word frequency also needs that qualitative interpretation depending on context.

Theoretical Aspects

The invitation to this workshop raised the question of the contribution which word frequency makes to linguistic theory. On the basis of the present evidence, I would suggest that the contribution is both *post-hoc* and *propter-hoc*. I've shown that frequencies are factors in items, systems, texts and discourses, that frequencies are discovered as part of distributional preference, that frequencies are used to indicate distributional choices, and that frequencies are quantitative but depend on qualitative interpretation. So I would suggest that frequencies are essentially calibrating – comparing but also establishing identity and discriminating individuality. Frequencies belong to description and prediction.

Conclusion: Use or Misuse?

In addressing my own question, I would conclude that 'misuse' is the statistical word. If all word frequencies were based on the statistical word test, nothing would follow or be revealing. All linguistic interest is in the frequency of the different types of words; as shown, frequency is a factor in the description of each type, not paradigmatic with the other types. There are only nine word types.

With regard to good use, I have shown that frequency is a factor with all word classes, that frequency is bound up with the interpretation of the value of the frequency of that word in the social context of occurrence, that frequency has a value in the description of particular lexical and grammatical items, and that frequency is replicable as a basis of systematic comparison and of identity construction. I conclude that it does not matter whether 'each text contains (approx.) 2,000 words' – rather it is the classification and interpretation of those 2,000 words in that particular text and context which will determine the real value of frequency study.

To go back to the beginning, I have shown that:

- Word frequency is the placing of numbers on language or the representation of language through numbers
- Word frequency provides an instantiation of the claim that 'linguistics is the scientific study of language'
- Word frequency promises precision and objectivity whereas the outcome tends to be imprecision and relativity
- Word frequency is not an end in itself but needs interpretation through contextualisation whence the relativity and comparative discrimination
- Word frequency is not science but methodology which lends itself to replicability.

Notes

[1] McArthur, Tom, 'What is a Word?' in *Oxford Companion to the English Language* (Oxford: OUP, 1992); Reprinted in McArthur, Tom, *Living Words: Language, Lexicography and the Knowledge Revolution* (Exeter: Exeter University Press, 1999).

[2] ICE-Ireland is the abbreviated name for The Irish component of the International Corpus of English. Cf. John M. Kirk, Jeffrey L. Kallen, Orla Lowry, Anne Rooney and Margaret Mannion, *The International Corpus of English: Ireland Component* [CD], Queen's University Belfast, ICE-Ireland Project, 2005.

[3] The British component of the *International Corpus of English*.

[4] Louw, Bill, 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies'. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.) *Text and Technology* (Philadelphia/Amsterdam: John Benjamins, 1993).

[5] *Northern Ireland Transcribed Corpus of Speech*, compiled by John M. Kirk, 1990, revised edition, 2004 (ESRC Data Archive, University of Essex).

[6] Kirk, John M., 'Aspects of the Grammar in a Corpus of Dramatic Texts in Scots', Ph.D. thesis (University of Sheffield, 1986).

[7] *The Miller Corpus of Undergraduate Conversation in Edinburgh Scots*.

[8] *A Corpus of Dramatic Texts from Glasgow*, compiled by John M. Kirk (Oxford Text Archive)

[9] Landau, Sidney, *Lexicography* (Cambridge: CUP, 1984).

[10] Cf. John M. Kirk, Jeffrey L. Kallen, Orla Lowry and Anne Rooney *The PPD Corpus* [CD], Queen's University Belfast, [beta version launched 2005].

[11] *The International Corpus of English* project comprises some 18 national varieties of standardised English. Those countries completed so far are East Africa (Tanzania, Kenya and Uganda), Fiji, Great Britain, Hong Kong, India, Ireland, New Zealand, the Phillipines, and Singapore. See <http://www.ucl.ac.uk/english-usage/projects/ice.htm>.

[12] *The London-Lund Corpus of Spoken British English*, see <http://khnt.hit.uib.no/icame/manuals/londlund/index.htm>

[13] *The Lancaster-Oslo/Bergen Corpus*, see <http://khnt.hit.uib.no/icame/manuals/lob/ index.htm>

[14] *The Freiburg Lancaster-Oslo/Bergen Corpus*, see <http://khnt.hit.uib.no/icame/manuals/flob/ index.htm>

[15] *Corpus of Dramatic Texts from Glasgow*, compiled by John M. Kirk (Oxford Text Archive).

[16] *The Leuven Theatre Corpus of British Dramatic Texts, 1966–72*, see Geens, D., L.K. Engels, and W. Martin. *Leuven Drama Corpus and Frequency List* (Leuven: University of Leuven, Institute of Applied Linguistics, 1975).