The Role of Frequency Effects in Linguistic Theory

Paul Bennett

1 Preliminary

Three positions on frequency in language:

A. It is a phenomenon that needs to be explained

B. It can be used to explain other aspects of language

C. It is an epiphenomenon of no theoretical importance

‘About half of the vocabulary of a text – even a very long text – consists of words that have occurred only once in that text’ (Sinclair 1991, p. 18).

Two examples from Newmeyer (2003). (1) Intransitive walk is more frequent than transitive walk. Does this require probabilistic information to be associated with subcategorisation frames, or is it due to the non-linguistic fact that walking oneself is more common than walking some other creature?

(2) Sentential subjects are far less common than extraposed structures (as in It is ADJ that S). Is this part of our knowledge of grammar, or due to processing reasons?

Frequency is a matter of number of occurrences in a corpus. Probabilities provide a way of modelling frequency (including combinations of events). Probabilities can also capture gradience (which contrasts with categoricity).

2 The Orthodox Generative View: Frequency is Irrelevant

‘the notion “grammatical in English” cannot be identified in any way with the notion “high order of statistical approximation to English”’ (Chomsky 1957, pp. 15–16). Neither Colorless green ideas sleep furiously nor Furiously sleep ideas green colorless has ever occurred in natural discourse, so statistically both are remote from English, but the former is grammatical and the latter is not.

In spoken English, adjectives are far rarer in pre-nominal than in predicative position, and especially rare in definite NPs. However, ‘any adequate grammar of English must provide for the possibility of adjectives occurring in definite NPs. Since, apparently, such adjectives are scarce in (some) discourse corpora, then that only serves to demonstrate that discourse corpora are insufficient as data bases for the study of the mental
representation of language. . . . generalizations about language structure should not be confused with the explanation of why certain forms are more likely to occur than others in a particular assemblage of texts.’ (Newmeyer 1998, p. 42)

‘probabilistic information drawn from corpora is of the utmost value for many aspects of linguistic inquiry. But it is all but useless for providing insights into the grammar of any individual speaker’ (Newmeyer 2003, p. 198).

Newmeyer (1998, p. 134): language learners and users are sensitive to frequency of grammatical elements, but that does not make frequency a legitimate explanatory factor.

3 Usage-Based Models: Frequency is Important

‘Since frequency of a particular usage pattern is both a result and a shaping force of the system, frequency has an indispensable role in any explanatory account of language’ (Barlow & Kemmer 2000, p. x).

Frequency is clearly relevant in applied areas such as lexicography and language teaching. And there are recognised frequency effects in psycholinguistics (e.g. more frequent words are recognised more quickly than less frequent ones). Much sociolinguistic research deals with the frequency of different forms of a variable.

Frequently-occurring items undergo some sound changes first (though this is controversial). Changes in the frequency of syntactic structures can lead to reanalysis (resetting of parameters).

Hare, Ford & Marslen-Wilson (2001): frequency of past-tense forms affects processing in both irregular and regular verbs. Subjects heard items that were ambiguous between past tense and a monomorphemic homophone (e.g. allowed/aloud). Where past tense was at least twice as frequent as homophone reading, subjects most often wrote down the past tense form.

Jurafsky, Bell, Gregory & Raymond (2001): the more probable a word (determined by its own frequency or its predictability in terms of the surrounding words), the more likely it is to be reduced in speech.

Hay (2001): relative, not absolute, frequency, is what matters: derived words that are more frequent than their bases are seen as less complex than those less frequent than their bases (cf. sane/insane and firm/infirm).

Tomasello (2000) argues that children have difficulties going beyond what they hear when they’ve heard it several times and so it’s entrenched. In an experiment aimed at inducing them to overgeneralise, children were more likely to produce I arrived it then I comed it. Note that arrive is less strongly entrenched than come, i.e. come is learned earlier by children and used more often by adults. ‘This finding suggests not only that children say what they hear, but that the more they hear it the more it seems to them that this is the only way it can be said’ (p. 72).
4 Some Recent Work Involving Frequency

Haspelmath (2006): frequency of use makes several senses of markedness superfluous. A general principle of least effort results in frequent signs being shorter. Frequency of use also explains markedness reversal, where categories that are usually unmarked are marked (cf. Welsh plu-en ‘feather’, plu-∅ ‘feathers’): nouns that are commoner in the plural have uncoded plurals.

Newmeyer (2003): 3rd-person reflexives are textually far commoner (in English, at least) than 1st- and 2nd-person reflexives. From this and the view that more frequently-used concepts are more likely to be lexicalised follows the universal that if a language has 1st- and 2nd-person reflexives, it has 3rd-person reflexives.

Adger (2006) examines the Scots dialect of Buckie, where there is variation we was/were and you (ones) was/were. With 1st-person plural 67% of instances have was. The feature system that Adger sets up involves we as [singular:–, participant:+, author:+]. There are two ways of deriving was: [uparticipant:+] and [uauthor:+] (where u = ‘uninterpretable’) and only one way of deriving were: [usingular:–]. So one would expect roughly twice as many instances of was as were. (See also Hudson (2007) and Adger (2007).)

Pierrehumbert (2003): English medial cluster such as /lskr/ are unattested. But this is because its subparts are rare, and a grammar that includes positional probabilities for phonemes will predict its absence.

Manning (2003) argues that many subcategorisation frames claimed by linguists to be ill-formed (e.g. regard + NP + to-infinitive) do occur, but very rarely (less than 1 in 100 occurrences of the verb in question). So syntax is not categorical (an either–or matter). Items in -ing can move gradually from being participles to prepositions; words move ‘in a continuous space of syntactic category, with dense groupings corresponding to traditional parts of speech’ (p. 315). A constraint that passives are marked will be categorical in some languages (so they won’t have passives) and soft in others (so passives will be less frequent than actives).

Bresnan (2006, p. 17): ‘language users’ implicit knowledge of their language is more powerful than has been recognized under the idealizations of categorical models of grammaticality: language users can in effect make accurate probabilistic predictions of the syntactic choices of others.’

Within computational linguistics, there has over the last 15–20 years been a renewed interest in frequency and probabilistic approaches. These use notions such as the probability of phrase-structure rules (i.e. of different ways of re-writing a category such as S) or of a particular lexical item being used in a particular way. A language model covers the likelihood of particular word sequences. Pereira (2000) proposes a statistical model based on bigrams (sequences of two words) and claims that Colorless green ideas sleep furiously is around 200,000 times more probable than Furiously sleep ideas green colorless.
References

Newmeyer, F. (2003), Grammar is grammar and usage is usage, *Language* 79, 682–707. (See also various responses in *Language* 81 (2005)).
Yang, C. (2008), The great number crunch, *Journal of Linguistics* 44, 205–228. (Review article on Bod et al. (2003)).