Frequency and sequence: highest, first – lowest, last

Michael Barlow (University of Auckland, New Zealand)

The traditional approaches to the ordering of linguistic elements can be summed up as follows. The discourse perspective on sequencing is typically treated in terms such as given-new or topic-comment. Written or spoken output is about something and the sequencing has been seen in terms of a topic or theme, which typically occurs in initial position, and a following comment or rheme. In spoken discourse, there may be an expression in initial position that acknowledges the previous contribution. In addition, for each language there is a set of conventions or syntactic constraints that determine the kinds of structures permitted and hence ordering. Thus in English, the default order is Subject-Verb-Object and there are constraints of various kinds: adjectives precede nouns, and so on.

The influence of discourse and grammatical constraints is clearcut. In this paper we investigate ordering from another perspective. We can assume that there is a trend for high accessibility words to occur before low accessibility words but quantifying accessibility is a difficult task and so in this research we focus on frequency. Using a corpus of newspaper articles and a corpus of spoken American in a professional setting, we examine the frequency of words in different positions in sentences/utterances. That is, we investigate the frequency or rank of a word in first position in a sequence, compared with second position, and so on.

If we ignore context for the moment, we might expect a sequence to start with high frequency words and progress to lower frequency words as the sequence progresses. The higher frequency of the initial elements means that they are more accessible for the hearer (and also the speaker). We can think of this as a lexical accessibility.

There are different approaches to determining the frequency of each word in a sequence. It would be possible to use raw frequency but this measure will give an exaggerated view of the distance between words. An alternative is to use rank or log rank as the starting point since this reduces the variation between words of different frequency and so smooths out some of the variation.

The next question concerns which frequency list to use. An initial thought might be to use the BNC as a large corpus with general coverage. However, the generality of the corpus is not so essential and may be problematic when dealing with a single genre such as news stories from a single newspaper. Therefore the frequency list used is taken from the corpus being investigated and it seems reasonable to assess the frequency distribution of words within sentences by looking at their occurrence in the corpus as a whole. Thus, a frequency list is generated for the corpus and from that the rank for each word is used as an indicator of accessibility. Hapax legomena are not included in the frequency list.

The two corpora used are a Times newspaper corpus and a selection of files related to committee meetings from the Corpus of Spoken Professional American English The latter consists of transcripts created by professional transcribers rather than linguists and so contains little in the way of prosodic features. The spoken usage is represented using sentences rather than prosodic units though with interruptions and false starts indicated by "—".

Since the number of words in a sequence varies, it is necessary to extract sentences/utterances of a set length so that we can get an overall picture of the frequency distribution in sequences of different lengths. For the spoken corpus, the sequences range from two words to ten words and for the written corpus, the ranges is 8- to 36-word sentences. The rank of each word in the sequence is provided. (The data is extracted and processed using python scripts.) Since turn-initial and non-turn-initial utterances may differ in their frequency profile, they are distinguished. An example of an 8-word turn-initial utterance is *I don't know if it's possible or not.* The rankings for each word in the sequence are: 6, 43, 39, 29, 28, 271, 33, 24. A non-turn initial sequence is: *the whole licensing mechanism is still under discussion,* which has the rankings: 1, 183, 524, 559, 8, 202, 294, 197.

Since the positioning of very frequent words such as articles may bias the results, the data is also processed with the most frequent words omitted.

If we examine the two sets of rankings above, we see that it is not the case that high frequency words always precede low frequency words. We do find in these instances, that the first word is the most frequent. It is to be expected that there will be a lot of variation in individual sequences but we wish to get a view of the general trends and do this we take the median rank value for each position in each sequence (8 words, 10 words etc.).

The results are remarkably consistent overall. For all the sequences examined in both the written and spoken corpora, we find that using a median value, the first word is always the most frequent and the last word is always the least frequent. This is the major finding. There are some more subtle variations in the frequency trajectory in different types of sequence that can be explored further.

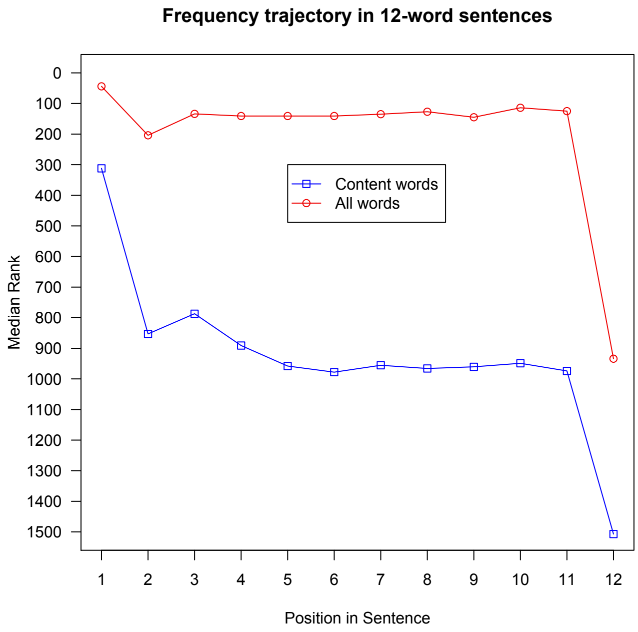

Figure 1: Rank of words at each position in a 12-word sentence

The graph in Figure 1 provides an example of these findings. We see clearly the first word effect and last word effect, which holds even if the most frequent words are omitted. We also see that beyond these boundary points, there are patterns in the frequency trajectory. We do not get a straight line from high to low frequency. We can examine these patterns for both the written and spoken corpora and for sequences of different length.

As noted above, there are well-known constraints on the ordering linguistic elements. English has fairly strict grammatical word order constraints and there are discourse processes relating to information flow. Despite these constraints, the results described here show that there are also some frequency effects, which can be seen when viewing the median rank values for sequences of the same length.

References

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics* 27(3), 517–56.