Stubbing your toe against a hard mass of facts: corpus data and the phraseology of STUB and TOE

Hans Lindquist
University of Växjö

Abstract

In this paper Fletcher's database Phrases in English is used to extract frequently recurring n-grams containing the verb 'stub' and the noun 'toe' from The British National Corpus. After analysing some of these n-grams, the paper focuses on the formulaic sequence 'stub one's toe' and investigates this in the BNC, The New York Times and The Independent. Additional searches are also made on the World Wide Web by means of WebCorp. It is found that the phrase is used with equal frequency in American and British English, but that the American use differs in that approximately half of the tokens are non-literal, while such use is relatively rare in British English. It is hypothesized that the non-literal use originated in American English and that it may be spreading to other varieties.

Somewhere down there he stubbed himself against an ill-defined but hard mass of fact, and brought it up to the surface to examine it.
(Michael Frayn: Towards the end of the morning, 1967; BNC G12 1612)

1. Introduction

Stubbs (1996, 2001) pioneered the use of corpora in the study of semantic and pragmatic meaning. In a number of more recent articles (Stubbs 2002, in press, forthcoming a, forthcoming b), he has suggested methods for retrieving n-grams from corpora in order to study frequent collocations and collocations with frequent words. These methods constitute innovative ways of bringing data, hard masses of fact, from the depths of large corpora to the surface where they can be examined and analysed.

The present study will use some of the methodology suggested in these papers to investigate the phraseological patterns, or, with Wray's (2002) terminology, the formulaic sequences which form around two lemmas, the fairly frequent body-part noun TOE and the rather infrequent verb STUB. My purpose is first to describe and analyse the semantic and pragmatic features of recurring formulaic sequences with these words, and, second, to discuss some theoretical and methodological consequences of this type of study.

The paper will have the following structure. After a brief section on the method and material there will be a few paragraphs on the history of the individual words stub and toe. Then the phraseological tendencies of these two words will be illustrated by a study of their occurrences in formulaic sequences in
the British National Corpus, which leads up to an investigation of the specific phrase *to stub one's toe* in the BNC, *The New York Times, The Independent* and on the World Wide Web (by means of WebCorp). Finally there is a conclusion and a Coda.

2. Method and material

The method used has been called "from lexis to n-grams" by Stubbs (forthcoming a) and is described in some detail in Lindquist and Levin (forthcoming a and b). Basically it means starting with a particular word or lemma, or a set of words or lemmas, and investigating which recurring n-grams they occur in. Lists of n-grams in the British National Corpus with the search word in all possible positions can easily be extracted by means of William Fletcher's (2003/2004) database *Phrases in English*, which includes all n-grams between 2 and 8 words occurring 3 times or more in the BNC.

At the next stage, these recurring n-grams have to be manually analysed to judge which may be considered to be formulaic sequences and which may be just chance occurrences without interior structure and integrity. For instance, in the BNC, the most frequent 3-gram beginning with *toe is toe of his* (21 occurrences), but this is not a likely candidate for a formulaic sequence to be stored holistically in the brain. The second most frequent 3-gram beginning with *toe*, however, is *toe the line* (16) which is clearly a formulaic sequence.

In the present paper, the phraseologies of *stub* and *toe* were also studied in two sets of newspapers on CD-ROM: the *New York Times* and the *Independent* from 1990, 1995 and 2000. Here, searches for the words were made by means of the program Wordsmith to create concordances which were then sorted and analysed manually. Finally, searches were made on the Web through the mediation of WebCorp (2007). However, as has been shown by e.g. Mair (2006) and Lüdeling et al. (2007), mining the rich resources of the Web is complicated due to a number of technical obstacles caused by the conflict between the information the linguist wants to extract through for instance WebCorp and the information that search engines like Google provide. Furthermore, due to various technical limitations, WebCorp at present returns many fewer hits than direct searches through Google. It is not the aim of the present paper to specifically evaluate WebCorp or Google data in comparison with data retrieved from traditional, tidy corpora or text archives like newspaper CD-ROMs. The Web data has however been added as a complement to the findings based on the other corpora. As has been pointed out by Mair (2006: 370), "corpus linguists of the future will […] [be] working in a vast and expanding corpus-linguistic working environment in which one of the chief skills required will be to identify the resources which are relevant to the problem studied from a vast range of possibilities". Comparing data from different corpora will often be a necessity (cf. Lindquist and Levin 2000).