Pre-publication version, November 2010.

Published in J Mukherjee & M Huber eds (2012) *Corpus Linguistics and Variation in English: Theory and Description*. Amsterdam: Rodopi. 223-30.

CORPORA AND TEXTS: LEXIS AND TEXT STRUCTURE

Michael Stubbs

University of Trier, Germany

ABSTRACT

Corpus studies have shown that words in texts have two strong tendencies: they occur in clusters and they occur in partly fixed phrases. These features have text-management and evaluative functions, and therefore provide the basis of a theory of how lexis contributes to textual organization. Although there are many excellent individual case studies of these topics, they have not yet been systematically integrated into a functional theory of lexis.

1. INTRODUCTION

In the last few years an increasing number of case studies have used corpus data to study textual organization, including the characteristics of specific genres (e.g. academic and legal) and their different conventions (Aijmer & Stenström eds 2004, Connor & Upton eds 2004, Partington et al eds 2004, Baker 2006, Biber et al eds 2007, Ädel & Reppen eds 2008, Adolphs 2008, Flowerdew, L. 2008). It follows that none of the facts which I discuss are very original, but I will argue that much work requires to be done before these case studies can be synthesized and integrated in a way which would reveal their full significance for a general theory of language.

I begin with a simple point. Corpora are artificial sets of data, which are produced by linguists, whereas texts are real language events, which are produced by language users. So, it is really texts – spoken and written – which should be the focus of attention. Corpora are one or two levels of abstraction away from real language events. In particular, concordances, one of the main tools of corpus study, encourage the analysis of small fragments of language, which are extracted from a cross-section of texts.

I will divide my examples into three main sets, which go from more concrete to more abstract:

1. word distribution: word-forms and semantically related words
2. word classes: what have been called general nouns and shell nouns
3. phraseology: particularly extended lexical units.

2. WORD DISTRIBUTION

I will start with the contribution of word distribution to text structure, and give just one example of each of two cases: the distribution of mere word-forms, where the analysis can be automatic and knowledge-free, and then the distribution of words in semantic fields, where the analysis depends on semantic interpretation.

The concept of knowledge-free text analysis famously goes back to work by Zellig Harris (1952). Although his specific proposals have rarely been directly developed, frequency and distribution are the features *par excellence* studied in corpus linguistics, and some text analysis makes either very simple assumptions about common-sense knowledge of the world or none at all. In an early quantitative analysis of word distribution in texts, Alford (1971) studied content words in Russian texts about physics: He pointed to an important difference between general high-frequency vocabulary and local high-frequency vocabulary, calculating that "once a general low-frequency word has occurred in a text, its immediate future text-coverage is likely to be more than ten times that predicted by [a] general count".

Church and Mercer (1994) make a similar point about the "burstiness" of lexical distribution. They note that content words in texts behave like buses in New York. You can wait for ages for a bus and none arrives, then several arrive together. I've never been to New York, so I can't confirm this, but I can corroborate the generalization from independent observations of buses in London: it may be an omnibus universal. Manning and Schütze (1999) put it more formally: "Most content words are much more likely to occur again in a text once they have occurred once". Kilgarriff (1997) calls it the "whelks problem". The lemma WHELK occurs over 70 times in the 100 million words of the BNC (British National Corpus), but around 40 of the occurrences are from one single article. In terms of text analysis, of course, it is not a problem, but a predictable and probabilistic feature of lexical cohesion.

One way of operationalizing a knowledge-free text analysis has been proposed by Youmans (1991). He calculates the type-token ratio in texts within a moving span of word-tokens: e.g. 1 to 35, 2 to 36, etc. For each new span, the software identifies "old" words (that is, words which have already occurred in the text) and "new" words (words occurring for the first time in the text). The changing type-token ratio is one indication of text structure. Figure 1 is an illustrative example from Joseph Conrad's short novel *Heart of Darkness*. This shows the first half of the text, with the type-token ratio measured over a moving span. The ratio tends to decline over the text as a whole, as more and more words are repeated, but the ratio rises at points in the story when something new is happening: new characters are introduced or there is some new incident. For example, there are clear spikes in the ratio at points A and B.

FIGURE 1 ABOUT HERE.

The spikes have been found with no prior assumptions about meaning, but of course their significance requires human interpretation. If we go back to the text,

we discover that something new is, indeed, happening at points A and B. At A, a group of people arrive, and the main character, Marlow, clearly disapproves of them: they are “an invasion, an infliction, a visitation”. At B, Marlow makes an “extraordinary” and “amazing” find, an old book in the middle of the jungle. If you have read *Heart of Darkness*, you will know that this is an important symbol in the story. The software has successfully identified significant points in the text. The “new” words are here marked with “> ”.

Point A:

> ... It was an >inextricable >mess of things decent in >themselves but that human folly made look like >spoils of >thieving. This devoted band called itself the >Eldorado >Exploring >Expedition and I believe they were >sworn to >secrecy. Their talk however was the talk of sordid >buccaneers, it was >reckless without >hardihood, >greedy without >audacity, and >cruel without >courage; there was not an >atom of >foresight or of >serious intention in the whole >batch of them ...

Point B:

> ... I >picked up a >book. ... It was an extraordinary find. Its >title was *An >Inquiry into some >Points of >Seamanship* by a man, >Tower, >Towson - some such name - >Master in his >Majesty’s >Navy. ... I >handled this amazing >antiquity with the greatest >possible >tenderness, >lest it should >dissolve in my hands ... Not a very >enthralling book; but at the first glance you could see there .. an >honest concern for the right way of going to work ... I couldn’t believe my eyes. ... It was an >extravagant mystery. ...

Other studies show that the uneven distribution of words from given semantic fields signals text structure. Here the analysis is clearly not knowledge-free, but depends on a semantic interpretation of the words. For example, Bondi (2007) studies economics articles and shows that different words are preferred in opening paragraphs (examples in 1 and 2 below) and in the main body of the text (examples in 3):

1. debate, economics, research, studies, theory
2. developed, modern, popular, rapid, seminal
3. coefficient, equation, increases, negative, parameter(s), rate, regression, total

In the opening paragraphs of such articles, there are frequently references to the discipline (often nouns) and evaluative words (often adjectives), whereas in the main body of the text there are statistical and logical expressions. As is often the case, this may seem rather obvious in retrospect, but such findings are not evident to introspection.

3. WORD CLASSES

Many studies have shown the contribution of word classes to textual cohesion. I will give brief examples of what have been called general nouns and shell nouns.

The top nouns in the BNC by descending frequency are :

> time, people, way, years, year, work, government, £#, day, man, world, life, part, mr, number, house, children, system, case, place, end, group, things

These words are not a random set. A couple (e.g. *government*, *£#*, where # represents any number) imply something about the design of the BNC (possible over-representation of topics common to newspaper texts). However, several have to do with time, place and people, and one or two have very vague meanings (*things*). That is, almost all are "general nouns", in the sense of Mahlberg (2005). In a recent article (Stubbs 2007), I studied why one of these words, *world*, is so frequent. The answer is that it occurs in many semi-fixed phrases.

1. from all over the world; in many parts of the world; the other side of the world
2. World War, Second World War, World Cup, Third World, World Bank
3. one of the world's most gifted scientists; the most natural thing in the world

Those in 1 refer literally to the external world: they are idiomatic and conventional but semantically transparent. Those in 2 are not entirely compositional, and those in 3 are evaluative phrasal constructions which have the function of emphasising something in the text.

In summary: High frequency nouns are frequent because they occur in frequent phrases, and many of these phrases are frequent because they structure and evaluate information. In addition, as Mukherjee (2007: 143) points out, high frequency is usually taken as a criterion of typicality. But the most frequent nouns in the language are not typical of the class of nouns: they do not denote categories of things in the external world, but have text-internal functions. In order to explain lexical frequency data, we require to show how lexis is used in texts.

Several studies go a step further and classify words functionally, according to their metalinguistic contribution to text structure: they encapsulate, interpret and evaluate a piece of information in the text. Many different terms have been used to label such lexis: vocabulary 3 (Winter 1977), procedural vocabulary (Widdowson 1983), prospective vocabulary (Tadros 1994), shell nouns (Francis 1994, Schmid 2000), signalling nouns (Flowerdew, J. 2006), unspecific anaphoric nouns (Yamasaki 2008), and probably others. Shell nouns refer to something spoken or written (1 below), or to thought processes (2 below), and encapsulate a piece of information in the text, and thereby make a metalinguistic contribution to text structure (Hunston & Francis 2000). Examples are:

1. accusation, boast, claim, demand, excuse, forecast, guess, hint, implication, etc
2. agreement, belief, concept, desire, fear, hunch, illusion, judgement, etc

Mahlberg (2009) gives an example which relates a genre-specific shell noun to text structure. In news stories, the noun *move* summarizes previous text, and the phrase *The move follows* often occurs at the beginning of the second paragraph of news stories. In the BNC there are 50 examples, all from spoken or written news media. Here is one example from BNC file CBF (periodical, world affairs):

> The BBC is to be stripped of responsibility for the Queen's annual Christmas TV broadcast ... The move follows ... claims that a BBC employee leaked the speech to a newspaper.

Hoey (2005) calls this tendency of certain words and phrases to occur at particular positions in texts "textual collocation" and "textual colligation".

4. PHRASEOLOGY

Finally, there are also many case studies of the text management functions of extended lexical units. In informal discourse, several constructions are used both to close a segment of narrative and to signal speaker attitude. The END UP VERB-ING construction (Louw 2000: 51) expresses irritation at a situation which leaves someone's image or status impaired (*you may end up looking a bit foolish; it ended up costing a fortune; we might end up regretting it*). The END UP IN construction signals a boundary in a story, often about something which started well but has gone wrong (*landing downwind and ending up in a hedge*). The most frequent collocates are *prison, jail, court* and *hospital*. It is not surprising that the lemma END occurs in phraseology which signals a narrative boundary, but the phraseology is conventional, not entirely compositional, and signals speaker attitude. The PAR FOR THE COURSE construction (Channell 2000) is a way of complaining that things have turned out disappointingly, but in rather the way you expected, and that there is nothing further to be said (*it didn't take [her] long to realize that this sort of thing was par for the course ... and that you have to learn to live with it*).

The WENT AND VERB-ED construction also signals the end of a segment of narrative, and often in addition expresses the speaker's surprise and/or disapproval of what has happened.

1. he put the phone down and went and got himself a whisky
2. so I went and toddled off to find somebody
3. then he went and jumped out of a plane
4. Paul Bodin then went and missed a penalty
5. then she went and spoiled everything by behaving as if pissed
6. and then, would you believe it, she went and married him

Again, it is important to relate such observations back to texts. Example 6 is from BNC file BN6 (non-academic prose and biography).

> ... a sad and curious story ... it concerns my great-aunt [and] a man ... called John Bell ... he was as mad as a hatter, and really no good at all ... and then, would you believe it, she went and married him ... the match was a disaster ... he made a lot of trouble ... and the family wanted nothing to do with him ...

5. CONCLUSIONS

I have used a few individual examples to make just one main point. Although there are many valuable individual case studies of the textual functions of lexis, they would contribute more convincingly to an overall functional theory of language use, if they were better synthesized and integrated.

There are many attempts to list and classify discourse markers of various kinds, but these taxonomies will never fully explain textual cohesion. This is because text structure is signalled by general mechanisms, including the uneven distribution of word-forms and of words which have been semantically and functionally classified. In addition, much frequent phraseology signals textual boundaries and/or evaluates a section of text. Much about lexical frequency is explained by textual function. The most frequent nouns are not typical nouns: they have text-internal functions. High frequency words are frequent because they occur in frequent phrases, and many phrases are frequent because they structure and evaluate information. It follows that implicit coherence (which relies on extra-linguistic knowledge) has been over-estimated, whereas explicit cohesion has been correspondingly under-estimated.

If we ignore the functions of lexis in texts, we miss the opportunity to make interesting generalizations. If we invent isolated sentences, and speculate about what is possible in theory (a Chomskyan view of things), then lexis appears chaotic. If we study attested texts, and look at what is probable in practice (a Firthian view of things), then lexis is much more predictable. This point reverses the classic Saussurian assumption. If we look at lexis as part of the language system (*langue*), many of its functions remain obscure. If we look at lexis as part of language in use (*parole*), many of its functions become much clearer.

6. REFERENCES

Ädel, A. and R. Reppen (eds.) (2008), *Corpora and discourse*. Amsterdam: Benjamins.

Adolphs, S. (2008), *Corpus and context*. Amsterdam: Benjamins.

Aijmer, K. and A-B. Stenström (eds.) (2004), *Discourse patterns in spoken and written corpora*. Amsterdam: Benjamins.

Alford, M. H. T. (1971), 'Computer assistance in language learning', in: R A Wisbey (ed.) *The computer in literary and linguistic research*. Cambridge: Cambridge University Press. 77-86.

Baker, P. (2006), *Using corpora in discourse analysis*. London: Continuum.

Biber, D., U. Connor and T. Upton (eds.) (2007), *Discourse on the move*. Amsterdam: Benjamins.

Bondi, M. (2007), 'Historical research articles in English and Italian', in: M. B. Papi, G. Cappelli and S. Masi (eds.) *Lexical complexity*. Pisa: Pisa University Press. 65-83.
Channell, J. (2000), 'Corpus-based analysis of evaluative lexis', in: S. Hunston and G. Thompson (eds.) *Evaluation in text*. Oxford: Oxford University Press. 38-55.
Church, K. W. and R. L. Mercer (1994), 'Introduction', in: S. Armstrong (ed.) *Using large corpora*. Cambridge, Ma.: MIT Press. 1-24.
Connor, U. and T. A. Upton (eds.) (2004), *Discourse in the professions*. Amsterdam: Benjamins.
Flowerdew, J. (2006), 'Use of signalling nouns in a learner corpus', *International Journal of Corpus Linguistics*, 11, 3: 345-62.
Flowerdew, L. (2008), *Corpus-based analyses of the problem-solution pattern*. Amsterdam: Benjamins.
Francis, G. (1994), 'Labelling discourse', in: R. M. Coulthard (ed.) *Advances in written text analysis*. London: Routledge. 83-101.
Harris, Z. (1952), 'Discourse analysis', *Language*, 28, 1: 1-30.
Hoey, M. (2005): *Lexical Priming*. London: Routledge.
Hunston, S. and G. Francis (2000), *Pattern Grammar*. Amsterdam: Benjamins.
Kilgarriff, A. (1997), 'Putting frequencies in the dictionary', *International Journal of Lexicography*, 10(2):135-155
Louw, B. (2000), 'Contextual prosodic theory', in: C. Heffer, and H. Saunston (eds.) *Words in context*. CD-ROM, English Language Research, University of Birmingham.
Mahlberg, M. (2005), *English General Nouns*. Amsterdam: Benjamins.
Mahlberg, M. (2009), 'Local textual functions of *move* in newspaper story patterns', in: U. Römer and R. Schulze eds *Exploring the lexis-grammar interface*. Amsterdam: Benjamins. 265-87.
Manning, C. D. and H. Schütze (1999), *Foundations of natural language processing*. Cambridge, Ma.: MIT Press.
Mukherjee, J. (2007) 'Corpus linguistics and linguistic theory' [Review of Mahlberg 2005], *International Journal of Corpus Linguistics*, 12, 1: 131-47.
Partington, A., J. Morley and L. Haarman (2004), *Corpora and discourse*. Bern: Peter Lang.
Schmid, H-J. (2000), *English abstract nouns as conceptual shells*. Berlin: de Gruyter.
Sinclair, J (1998), 'The lexical item', in: E. Weigand (ed.) *Contrastive lexical semantics*. Amsterdam: Benjamins. 1-24.
Stubbs, M. (2007), 'Quantitative data on multi-word sequences in English', in: M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert *Text, discourse and corpora*. London: Continuum. 163-89.
Tadros, A. (1994), 'Predictive categories in expository text', in: R. M. Coulthard (ed.), *Advances in written text analysis*. London: Routledge. 69-82.
Widdowson, H. G. (1983), *Learning purpose and language use*. Oxford: Oxford University Press.
Winter, E. (1977), 'A clause relational approach to English texts', *Instructional Science*, 6, 1: 1-92.

Yamasaki, N. (2008), 'Collocations and colligations associated with discourse functions of unspecific anaphoric nouns', *International Journal of Corpus Linguistics*, 13, 1: 75-98.

Youmans, G. (1991), 'A new tool for discourse analysis: the vocabulary management profile', *Language*, 67, 4: 763-89.

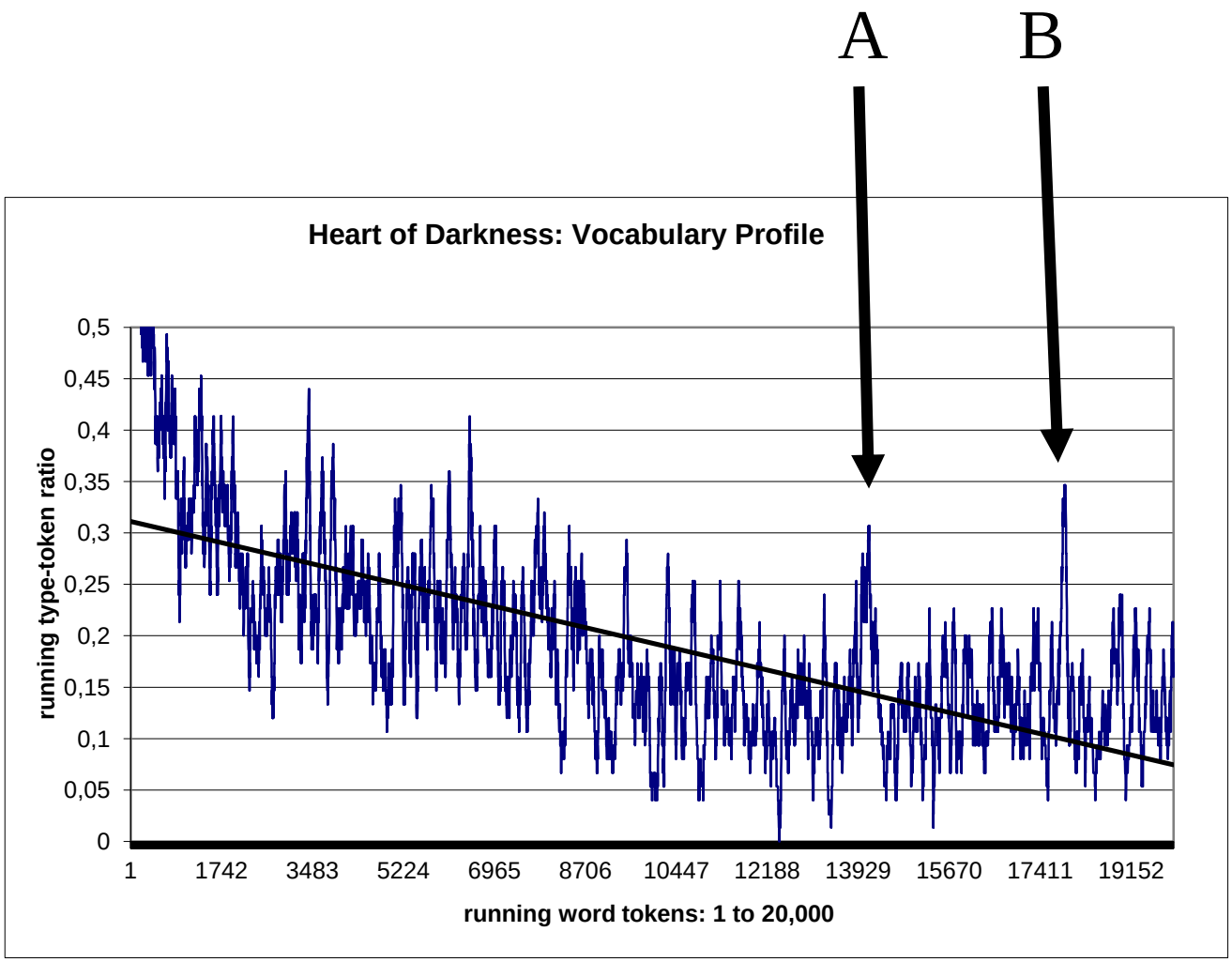

FIGURE 1. The moving type-token ratio across the first half of Conrad's novella.