Textual analysis

Why do it, and where does it take you?

Eirian Davies
Royal Holloway, University of London

Margaret Berry, in whose honour I am glad to contribute this squib, is, as far as I know, alone among practitioners of Systemic Functional Linguistics (SFL) in questioning one of its major articles of faith: namely, the over-riding importance of textual analysis for any valid statement about a language (e.g. Berry 1981: 61). This is a line of questioning that I believe can be fruitfully explored.

In my view, there are three fundamental reasons for studying the language of texts. These are to discover more about: (i) the natural language concerned, itself; (ii) the texts themselves, and/or those who produce them, and/or other features in the world mediated and conveyed by them; (iii) the theoretical model being used to account for them (see also Simon-Vandenbergen, this volume). I will confine myself to the third reason, as being perhaps the one least usually discussed.

There is a major difference in approaches within the area of studying texts for the purposes of developing linguistic theory. This is the contrast between deriving a theory from textual evidence (corpus-driven approaches) and testing a theory against corpora (corpus-based studies), a difference originally established by Tognini-Bonelli (2001). "Firthian" linguistics may be thought of as spanning both the corpus-driven approach, as in Sinclair's treatment of the study of lexis (Sinclair 2004), and the corpus-based approach, as in Halliday's work on grammar (e.g. Halliday & Matthiessen 2014).

Within corpus-based studies, I would suggest that there is also a largely undiscussed contrast between testing a linguistic theory, which tells us about the adequacy of the model, and testing a description of a given language, which tells us more about the language. A linguistic theory may be invalidated by being tested against the evidence provided by texts if it fails to account for material that exists in them. On that basis, we have learned something new about the theory and that theory will need to be revised or rejected. For example, if a linguistic theory offered no provision for distinguishing different parts of speech, we might want to argue that it failed for a language like Latin when tested out against texts.

https://doi.org/10.1075/fol.00015.dav
Functions of Language 26:1 (2019), pp. 43–48. issn 0929-998x | e-issn 1569-9765

On the other hand, in asking whether or not a description of a given language is adequate, we are not usually asking whether the categories of that description (derived from some pre-existing theory) are adequate as such, but, more, whether or not the language being described is fully accounted for by what the application of these categories has to say about it. The key point is the question of whether or not there is a category available in the descriptive apparatus which can account for the data at issue. If such a category does not exist, this implies that the theory from which the description derives is inadequate. However, if such a category does exist, but has not previously been thought to apply to a given language, then the description needs to be improved by including it. But, in that case, the theory which generates that category, and the other categories in terms of which the description is made, remains valid. For example, if we find tense markers in texts of a language not previously described as having them, we will revise the description, but not the categories in terms of which it is made. We will have discovered that there is more to the linguistic data than was previously known and we have learned something about the language.

Corpus-based approaches test out and modify categories already established in an existing theoretical model against the evidence available in large bodies of textual data; and they can also use corpora as the source of illustrations for such pre-established categories. This is the sort of approach endorsed by Halliday,[1] and generally applied in much of the work in different kinds of functional linguistics, including cognitive grammar, and work based on the Survey of English Usage at University College London. It is a mode of connecting with text that has been used as a corrective to some of the claims made in formalist models on the basis of native speaker intuitions alone. In this respect, it is based on what I would call counter-, as opposed to negative-, evidence. That is, it operates mainly with finding evidence of categories predicted not to exist, or predicted to exist in a different form, or with categories not predicted to exist, as opposed to cataloguing a lack of evidence for categories that are predicted.

For example, cases of counter-evidence would include the discovery in texts of evidence of a category such as the definite article (or the tense markers mentioned above) which had been predicted not to exist in a given language; or of inflected markers of modality in a language predicted to have modality conveyed only through modal auxiliary verbs; or the discovery of textual evidence of some category such as dual number which the theory underlying the description had not allowed for. All these cases would contrast with that case in which the underlying theory had predicted a category such as a particular kind of modality for

1. For an account of the kinds of use made in SFL of corpus evidence see Halliday & Matthiessen (2014: 69–74).

which no evidence has yet been found in texts. This is what I am calling the case of negative evidence. It can be compared with cases in the physical sciences in which a prediction is generated by a particular theory, for which there is no empirical evidence at the time the prediction is made.

Approaches involving the use of both positive and counter-evidence have made an important contribution to more adequate descriptions of a number of natural languages, although their contribution to theoretical developments is not always so clear cut. In this connection, we may contrast Halliday's position, when he claims that the analysis of text leads to theoretical insights (see Simon-Vandenbergen this volume), with that of the authors of one of the major corpus-based grammars, Quirk *et al.* (1985), whose approach is explicitly theoretically eclectic. The latter, together with much pedagogical work using corpora to inform and correct materials developed for the purposes of teaching English to non-native speakers (e.g. Downing 2015), can be said to be mainly oriented towards description.

However, we may question whether, even in the case of SFL, this form of testing out categories against texts has not typically been focused more on developing descriptive, rather than explanatory, adequacy: telling us more about the language than about the theoretical model. For example, it is not at all clear that insights into the major theoretical categories of the metafunctions in SFL have arisen from testing them out against textual material. Similarly, the concept of systems and their place in the structure of the theory cannot, I would argue, be shown to have been modified or adapted by being tested out against texts. The content of different systems has been often so modified; but this is modification of the description, as opposed to modification of the theory.

I believe that there is a strong case for arguing that, both historically and in more recent practice, the main motivation in SFL for the analysis of texts has been the second of the three reasons with which I began: that is, to tell us more about the texts themselves. This has probably been more pronounced and more successful than recourse to textual material for the purposes of developing the description. One of the best known instances of the application of SFL to the analysis of a text is Halliday's (1971) analysis of the language of William Goldings novel *The Inheritors*, as mentioned in both Simon-Vandenbergen's and Asp's contributions (this volume). This is directed at developing an insightful account of the meaningfulness of the text rather than at telling us either more about the English language as such or its description or about the nature of the theory that underlies the linguistic model used to describe the text. The theoretical status of the SFL metafunctions is not here modified by this application of a descriptive model founded on the theory of which they form part. We do not learn more about the theoretical status of the ideational metafunction from this analysis: we learn more about the way the Neanderthals are presented as seeing the world in this novel.

The other approach to using large data banks in developing a linguistic model is to be found within "corpus-driven" corpus linguistics. Here the model itself, and its categories, are constructed on the basis of the statistical analysis of large bodies of data. Models of this kind include the approach to lexis in the work of Sinclair and others at Birmingham,[2] and many studies of different kinds at Lancaster, Lund and elsewhere, including those by Biber and colleagues in the USA (e.g. Biber & Reppen 2015).

A distinction can be made between academic studies within corpus-driven corpus linguistics and other statistical work with "big data", which I use here as a shorthand label for a kind of statistical approach based on large bodies of data which is neutral to any specifically **linguistic** theory. That is, statistical approaches to big data banks may employ techniques of search and pattern recognition which would be equally well-adapted to the analysis of any other form of human (or non-human) activity. Academic corpus-driven linguistics remains concerned with linguistic theory (e.g. Hoey 2005; Hunston 2008; Moon 2009); but, I would suggest, the most striking advances made in it have occurred in applications rather than theory. Many of its major successes, which include the Cobuild dictionaries and the Longman grammars of English (based on the same corpus), have, in my view, been most notable for advancing the description of English, rather than for developing linguistic theory as such.

Both corpus-driven and big data approaches rely on what I shall call "positive evidence": that is, the presence of material in texts. Corpus-based approaches also explicitly take into account what I have called "counter-evidence" above: the presence of categories in texts which are predicted by a given theory not to exist. This can be used to invalidate a (pre-existing) theory, a possibility clearly not available in any approach in which the theory itself is directly derived from the analysis of whatever material actually does exist in a given data bank.

As mentioned above, there is, however, also a third kind of evidence that may be noted, which is the opposite of counter-evidence. I have called this "negative evidence": **absence** in the material studied of some category that **is** predicted by a (pre-existing) linguistic theory to exist. For example, the theoretical approach to epistemic modality in Davies (2006) allows for 16 categories with respect only to four basic sentence types and the mood of the lexical verb (together with putative *should (not)*). The number of categories predicted to apply to the whole area of epistemic modality, including those realized by the modal verbs in English, is substantially higher. This may be over-provision for English; but my argument is that absence of evidence of any one (or more than one) such category should not be taken as grounds for rejecting the theory. As with counter-evidence, negative evi-

2. For some coverage of this see the discussions in Moon (2009).

dence depends on a pre-existing linguistic theory, and cannot apply in a corpus-driven approach.

Care must be taken with negative evidence. With respect to testing out a linguistic description of a natural language, we cannot legitimately infer from the fact that any given construction type or lexical item does not occur in our data the conclusion that it does not occur in the language at all. With respect to testing out a theory, we cannot infer from the fact that a given category does not occur in the corpora of a given natural language the conclusion that it could not occur in that language. Further, we cannot infer from such negative evidence that a category does not, let alone could not, occur in some other natural language. That is, negative evidence on its own is not sufficient to disprove a theory.

The lack of occurrence in a data bank of a category predicted by a given theoretical model cannot be taken as evidence against that model. It might equally well be interpreted as evidence against the adequacy of the data bank. Especially in cases where native speaker intuition affirms the validity of a category, and can supply exemplification of it, there is a strong case for claiming that its failure to occur in a given data bank tells us more about the data bank than about that theoretical category. The possibility of being inadequate applies to data as much as to theory.

This is a general point about negative evidence and the validation of linguistic theories of any kind. If a theory is a universal theory of natural language, it cannot be automatically inferred from the fact that no illustration of a particular category which the theory predicts is found in a given natural language that the theory is therefore proven wrong. Such an absence of positive evidence for a category which that theory predicts may tell us more about that given natural language than about the theory. That is, by analogy, just as we may argue from the non-occurrence in a given corpus of a category predicted by the theory to a characterization of the corpus, so we may argue from the non-occurrence in a given entire language of a category established in the theory to a characterization of that language. An example of this would be the absence of the definite article in a language such as Korean. A universal theory must allow for categories which are found in some languages, but not in others. This is the reverse of the otherwise constricting claim that, to count as "universal" a model must consist in only those categories that occur in all languages.

The relationship between theory and data is not symmetrical. A theoretical model must be able to account for all categories and patterns that are found to exist, and fails if it cannot do so. Such failure tells us something about the model. But, if a theoretical model predicts categories that are not (yet) found to exist, that kind of failure may tell us something about either the model, or the data as a good representation of the language, or about the language itself. We may either test the

model against the data; or we may test the data against the model. Ideally, we need to do both; but I believe relatively little has so far been done on the latter.

With respect to the use of textual analysis as an aid in the development of linguistic theory, I suggest that much remains to be done. Perhaps an adjustment of focus which allowed for more theory-led investigation of data banks or corpora might invigorate the process.

References

Berry, Margaret. 1981. Polarity, ellipticity, elicitation and propositional development: Their relevance to the well-formedness of an exchange. *Nottingham Linguistic Circular* 10.

Biber, Douglas & Randi Reppen. 2015. *The Cambridge handbook of English corpus linguistics*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139764377

Davies, Eirian. 2006. Speaking, telling and assertion: Interrogatives and mood in English. *Functions of Language* 13(2). 151–196. https://doi.org/10.1075/fol.13.2.06dav

Downing, Angela. 2015[1992]. *English grammar: A university course*, 3rd edn. Abingdon, NY: Routledge.

Halliday, M. A. K. 1971. Linguistic function and literary style: An enquiry into the language of William Golding's *The Inheritors*. In Seymour Chapman (ed.), *Literary style: A symposium*, 330–369. Oxford: OUP.

Halliday, M. A. K. (revised by Christian M. I. M. Matthiessen). 2014. *Halliday's introduction to Functional grammar*, 4th edn. London: Routledge.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, Susan. 2008. Starting with small words: Patterns, meaningful units, and specialized discourses. *International Journal of Corpus Linguistics* 13(3). 271–295. https://doi.org/10.1075/ijcl.13.3.03hun

Moon, Rosamund (ed.). 2009. *Words, grammar, text: Revisiting the work of John Sinclair*. Amsterdam: Benjamins. https://doi.org/10.1075/bct.18

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Sinclair, John McH. 2004. *Trust the text: Language, corpus and discourse*. London: Routledge.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins. https://doi.org/10.1075/scl.6

Address for correspondence

Eirian Davies
Department of English
University of London
Royal Holloway
United Kingdom
e.davies@rhbnc.ac.uk