

*author*

**Michal Měchura**

*assignment title*

**Discourse analysis using corpora**

*degree*

**M.Phil. in Speech and Language Processing**

*module*

**LI 7864 Corpus Linguistics**

*term*

**Michaelmas 2007**

*word count*

**3,839 words**

# Discourse analysis using corpora

Michal Měchura

This essay investigates the current state of the art in corpus-based discourse analysis. It first gives a definition of what is meant by discourse and an overview is presented of the various discourse-related phenomena that have been investigated by corpus linguistics. Then, the main part of the essay deals in some detail with one particular discourse phenomenon, that of *information structure*. The investigation of information structure using corpora is illustrated on two corpus-analysis projects whose methods and results are evaluated and compared. Finally, the essay discusses the future directions in which corpus-based discourse analysis is likely to develop.

## 1. What is discourse analysis

Definitions of discourse analysis encountered in the literature typically afford a generously wide scope to the discipline. According to one introductory book, discourse analysis is an “enquiry into how people make meaning” (Widdowson 2007, p. XV). By that definition, the interests of discourse analysts should include matters such as morphology and syntax, because morphology and syntax are, after all, used to make meaning: *cat* and *cats* mean different things, and so do *dog bites man* and *man bites dog*. But discourse analysis is not interested in morphemes or in subjects and objects. Such matters are already covered in other disciplines. Discourse analysis analyzes sentences in terms of their communicative purpose. More typically, it analyzes not sentences but texts.

A text is an instance of language in use which, once again, has communicative purpose. When we humans come across an instance of language in the real world, we recognize it as a text if and when we realize that it attempts to achieve something (even if simply to provide information) and that it refers to something specific. For example, the sentence *I can't find it* cannot be analyzed from the point of view of discourse analysis unless the context in which it was uttered is known. Only then are we in a position to reconstruct what the pronouns *I* and *it* refer to, and whether the whole sentence is a request for help in finding this *it*, whether it has been uttered as an excuse of some sort, or indeed whether it has been uttered

simply to provide information. On the other hand, a linguist who is not a discourse analyst does not typically need to know the context in which a sentence has been uttered: it can be analyzed from the point of view of morphology and syntax, and even some semantics, when taken in isolation. With all its emphasis on communicative purpose, discourse analysis can be best understood as an exercise in pragmatics.

Discourse analysis assumes that there is something called discourse which exists quite outside the text. Discourse encompasses the meaning of a text as the producer intended it, and the meaning as the receiver reconstructed it. The text is understood as merely a trace of the discourse. The text is the producer's best attempt to encode what she meant, and it is the basis from which the receiver reconstructs the meaning. In most cases, the two participants' understanding of the discourse is identical, but not always. It is quite possible, for example, that the receiver resolves the reference of *it* differently from the way the producer meant it, or that the receiver misunderstands *I can't find it* as a mere provision of information when in fact the producer meant it as a request for help. In most cases, however, participants seem to agree on their interpretations of the discourse they are engaged in. When observing other people's discourse, a discourse analyst treats discourse as something that is not present in the text explicitly but that he can be reconstructed from it – bearing in mind that he may not always reconstruct it exactly as the producer meant it or as the intended receiver understood it, but that he probably will most of the time.

## **2. Discourse analysis and corpora**

The process of reconstructing discourse from a text, be it a single text or a corpus of texts, can be easy or difficult to varying degrees, depending on the text type. Spoken texts and texts of similarly spontaneous nature, such as on-line conversations, are strongly dependent on their immediate context with the consequence that it may be difficult to reconstruct the purpose and reference of utterances if the observer is removed from the event. Written texts, on the other hand, especially those of the formal type, tend to be constructed deliberately as independent of context as possible, and their pragmatic meaning may be easier to reconstruct.

In either case, discourse analysis amounts to inferring from the text something that is not present in it explicitly. This is nothing new for the corpus linguist: almost all the information corpus linguists are interested in is not present in the text explicitly, and must be inferred. Even part-of-speech information and constituent structure is not encoded in a raw text explicitly (because a raw text is simply a sequence of tokens) and must be inferred by parsers, automatic or human. It can be seen, however, that pragmatic meaning is at a further remove from the text than, say, constituent structure. Inferring pragmatic meaning therefore requires higher powers of inference. It is also more open to interpretation and disagreement, and more difficult to automate.

There are essentially two ways a corpus can be used for discourse analysis. Firstly, a corpus may already include some discourse-related annotation, and this can be queried by the researcher or used as data for machine learning. Secondly, a corpus that does not include any discourse-related annotation (or does not include the kind of annotation the researcher needs) can still potentially be used to investigate discourse but the researcher must compensate for the lack of annotation by employing smart workarounds and querying techniques. For example, even if a corpus does not annotate *well* as a discourse marker when it functions as such, the researcher can still call up all occurrences of *well* and filter out manually those that are not discourse markers, and then analyze the remainder.

In a way, discourse features have been part of corpus linguistics from early on. Even at the level of morphological annotation, some tag sets include tags for “discourse items” such as *well*, in addition to the more conventional part-of-speech tags for nouns, verbs and the like. Corpus annotators have always been aware that some words are (sometimes) used in such a way that assigning conventional part-of-speech tags to them simply would not make sense. Perhaps an extreme example is the spoken part of the Susanne corpus which recognizes as many as 14 types of discourse tags such as apology (*pardon, sorry*), response (*ah, fine, good*), greeting (*hello*), expletive (*hell*), engager (*you know*) and others. The written part of the Susanne corpus contains only a single tag for all types of discourse items (Sampson 1995, p. 445-448). Generally, tagging for discourse at the token level is more common in speech corpora. For example, the Childe corpus of (spoken) child language contains a tag called “communicator” for discourse-functioning tokens.

The intermixing of discourse tags with part-of-speech tags is not without controversy. In the sentence *hell I could use a beer now, is hell* a noun or a discourse item? The answer is of course that it is both, and ideally both facts would be recorded in a corpus. Annotating the discourse function of items belongs to a different level of annotation than part-of-speech tagging, and some modern corpora have recognized that – for example the Prague Dependency Treebank, which will be dealt with in more detail later.

In corpora where annotation is subdivided into different levels, discourse annotation is generally to be found at the higher levels. One linguistic phenomenon that clearly belongs at a very high level of annotation is the analysis of utterances in terms of *speech acts*. Speech act is an important category in discourse studies because it indicates the very communicative purpose of an utterance: question, provision of information, expression of opinion, offer of encouragement, and so on. There is no shortage of theoretical descriptive frameworks for classifying utterances into different categories of speech acts, but to my knowledge, no major publicly available corpus contains speech-act annotation – with the exception of the Childes corpus whose annotation scheme offers the facility to tag spoken utterances for speech-act type.

One discourse-level phenomenon that has been receiving the attention of corpus annotators for some time is coreference. A typical text contains chains of text-internal references in which proforms refer to noun phrases (*I have lost my glasses and now I can't find them*), noun phrases refer to other noun phrases (*We visited Donegal and really loved the place*), even ellipsis that refers to previously mentioned items (*You can come if you want to [come]*). A human reader resolves such references effortlessly: upon encountering the pronoun *them* in the first example above, it becomes immediately obvious that it refers to *my glasses*. It is much more difficult for a computer, however, to resolve these text-internal references automatically, and there have been several attempts to create corpora annotated for coreference with the goal to use those corpora to understand the features of coreference better and to train computers to resolve coreferences automatically.

It is hopefully obvious from the treatment so far that discourse analysis is interested in many phenomena to which corpus linguistics is applicable – at least in principle – and each such phenomenon would deserve its own essay. I would like to

devote the remainder of **this** essay to one area of discourse analysis that has already begun to be investigated by corpus linguists to a considerable extent: the phenomenon of *information structure*.

### 3. Information structure

The *information structure* of a sentence refers to the ordering and patterning in which information is presented in the sentence. There are two major competing theoretical frameworks that attempt to describe the phenomenon: the Hallidayan school (Halliday 1994) on the one hand, and a family of descriptive frameworks rooted in the Prague School of Linguistics on the other hand. They both agree on one basic principle: a clause can typically be subdivided into two halves, a beginning called the *theme* and a remainder called the *rheme*. A theme is understood to be the “point of departure” (Halliday 1994, p. 37) for the clause, it is how the writer links the clause to the preceding text. An alternative (and partially equivalent) definition is that the theme is what the sentence is “about” and the rheme is what the sentence says about the theme (Hajičová and Sgall 2004). The pair of terms *theme/rheme* is well established in the study of information structure, but some frameworks prefer to use different terminology, such as *given/new*, *topic/focus*, *topic/comment* or even combinations of these pairs. In this essay, *theme* and *rheme* will be used to refer to all of these binary distinctions in a framework-independent sense, unless indicated otherwise.

I will now present and evaluate two projects in which corpus analysis techniques have been used to investigate information structure. The first is a relatively early attempt to analyze information structure “by hand” in a small unannotated corpus, the second is a major, recent corpus project involving explicit annotation for information structure. These two projects are representative of the two different approaches that have been taken in the history of corpus-based studies in general: one is based on a raw corpus where the researcher analyzes concordance lines manually, hoping to notice any common features; the other involves extensive annotation for the phenomenon under investigation (that is, information structure) and then letting any findings to “emerge” on their own, essentially by computing statistics from the annotations.

#### 4. Kenny's corpus study of information structure in Irish and English

As mentioned, it is possible to investigate theme and rheme in a corpus even if the corpus does not contain any annotation that would clearly demarcate the theme from the rheme. Dorothy Kenny of Dublin City University has carried out a corpus-based, cross-linguistic comparison of information structure in English and Irish (Kenny 1998) using a small raw-text corpus of English weather bulletins (2,223 tokens) and their Irish translations (2,234 words) as they appeared in television broadcasts. She set out to find out what kinds of information tend to be thematized in the two languages.

In the Hallidayan framework (Halliday 1994, chapter 3), theme is defined simultaneously in terms of its discourse function and in terms of its linear arrangement in the sentence. The discourse function of a theme is to be the “point of departure” (Halliday 1994, p. 37) for the sentence, and one is supposedly able to recognize it by the fact that it is at or very near the beginning of the sentence. Halliday's framework is very specific about the theme being always at the beginning of the sentence and allows only a small number of exceptions when the theme may be preceded by other items. This appears to work well for English, but in a verb-initial language like Irish, this would lead to the unsurprising observation that Irish sentences have an overwhelming tendency to thematized the verb – as Kenny's corpus analysis duly confirmed.

This has lead Kenny to reject the Hallidayan framework as inappropriate for the description of information structure in Irish, and repeat her corpus analysis using a framework from the Prague School tradition, namely that of Functional Sentence Perspective (FSP – Firbas 1992). This framework operates with the notion of *communicative dynamism*. For example, a sentence whose discourse purpose is to present the existence of some phenomenon can be subdivided into three elements: Setting, Presentation and Phenomenon. The communicative dynamism increases gradually from Setting to Presentation and from Presentation to Phenomenon:

Setting:            *There*  
Presentation:    *will be*  
Phenomenon:    *rain*  
Setting:            *in the north-west.*

Elements with low communicative dynamism (such as Setting and partially also the Presentation) are understood to be members of the theme, while elements with high communicative dynamism are members of the rheme. Notice that the elements can be discontinuous and do not need to be present in any particular order. They are defined purely by their discourse function, although in a free-word order language like Czech, they typically do come in an order of increasing communicative dynamism:

Setting:	<i>Na severozápadě</i>	[ <i>in the north-west</i> ]
Presentation:	<i>bude</i>	[ <i>will be</i> ]
Phenomenon:	<i>děšť.</i>	[ <i>rain</i> ]

The more rigid syntax of languages like English and Irish does not allow the linear ordering of elements in a sentence to copy their increasing communicative dynamism this closely, but that changes nothing on the fact that sentences can still be analyzed into discourse categories like Setting and Phenomenon.

Having analyzed her corpus in this way, Kenny was able to make some interesting conclusions as to the information-structuring strategies of Irish in comparison to English. One of her findings is that while English weather bulletins have a strong tendency to delete items of low communicative dynamism, especially when those are expressed by low-content words like *there* and *will be* (for example *rain in the north-west* instead of *there will be rain in the north-west*, i.e. the Presentation and part of the Setting are deleted), Irish weather bulletins only delete such items with extreme reluctance. The pattern *there will be...* is very uncommon in the English corpus, but its Irish translation equivalent *beidh ... ann* is very common in the Irish corpus. One can restate these findings as revealing a lack of willingness in Irish to delete thematic/redundant/repetitive elements, compared to English – even though the observation is limited to the sublanguage of weather bulletins. This finding, while retrospectively appearing to be consistent with intuition, would have been difficult to formulate precisely without the analysis of a body authentic texts, i.e. without the employment of corpus analysis techniques.

This important finding of Kenny's belongs to one category of corpus usage: corpora are used for linguistic research to discover facts about a language which were not known before, or to validate pre-existing observations about a language. But there is another category of corpus usage, and that is the validation of

descriptive frameworks. Kenny has demonstrated that Halliday's framework of theme/rheme is (arguably<sup>1</sup>) unsuitable to the description of information structure in a verb-initial language like Irish, and that Firbas's Functional Sentence Perspective is a better attempt at devising a language-independent descriptive framework. Such findings may well be of greater importance than any language-specific observations.

### **5. Information structure in the Prague Dependency Treebank**

Kenny's analysis was performed on a raw corpus that contained no explicit annotation for theme and rheme. It can be expected that a corpus that does include such annotation will yield even more revealing observations. One such corpus is the Prague Dependency Treebank (PDT), a large corpus (1.8 million tokens) of Czech newspaper texts. The corpus is annotated at several "layers" which one can visualize as existing "above" the layer of the tokens themselves. The first layer (the *m*-layer) is the layer of morphological annotation (lemma and part of speech), the second layer (the *a*-layer or *analytical* layer) is where the syntax of sentences is annotated by way of dependency trees, and the top-most layer (the *t*-layer or *tectogrammatical* layer) contains an interpretation of the sentences in a predicate-and-argument structure with the roles of predicate arguments indicated in terms of functional categories such as *actor*, *patient* and *location*. The tectogrammatical layer is also annotated for several other phenomena including semantic number, semantic gender, politeness, modality, coreference and, importantly, information structure.

In many ways, the tectogrammatical layer in the PDT can be understood as a representation of some aspects of the discourse that underlies the text. It is therefore appropriate that this annotation is at a considerable remove from the actual text: the dependency trees in the tectogrammatical layer do not contain

---

<sup>1</sup> It must be mentioned that Halliday's functional grammar includes an additional pair of functions, *given/new*, to complement *theme/rheme* in its description of discourse functions. While *theme/rheme* is defined by Halliday with partial reference to the linear sequencing of elements in text, *given/new* is defined purely with reference to discourse. Combined together, the two distinctions aim to describe the same kind of phenomena as the various Prague School frameworks do. The *given/new* distinction was not used in Kenny's analysis, but had it been, it might have not resulted in her rejection of Halliday's framework.

nodes for all tokens. Some tokens – mostly function words – are not represented in this layer as separate nodes but merely as attributes of other nodes, while elements deleted in the token layer, such as ellipses, are explicated fully in the tectogrammatical layer. We can say that the information-structure annotation (as well as other types of annotation on this layer) is attached to data that has been **abstracted away** from the raw text. This is different from Dorothy Kenny’s analysis, which was done on the raw text itself.

The information structure in PDT is based on yet another descriptive framework in the Prague School tradition, namely Topic-Focus Articulation (TFA – Hajičová and Sgall 2004, Hajičová et al. 2003, p. 98–107). As might be expected, this framework divides a sentence into two parts, called here the *topic* and the *focus*. These two functions are defined conceptually in terms of “aboutness”: in a prototypical declarative sentence, the topic is what the sentence is about and the focus is what the sentence asserts about the topic. Typically, the topic precedes the focus, but this may not always be the case. Like in Firbas’s Functional Sentence Perspective, the two functions are defined purely with reference to discourse, without regard to their linear ordering at the token layer.

The concepts of topic and focus in TFA are related to the more precisely defined concepts of *contextual boundness* and *contrastivity*. An element in a sentence is *contextually bound* if it is presented in the text as though it refers to something already mentioned (even if it does not in fact corefer to anything textually – Hajičová et al. 2003, p. 100) or to something in the context of the utterance. Contextually bound items are prototypically (but not always) part of the topic. All other items are *contextually unbound* and are prototypically part of the focus. Additionally, an element can be *contrastive*, which simply means it is presented in the text as though it stands in contrast to something. Elements in the focus of a sentence are almost always contrastive because the focus almost always presents new information, and this new information can be understood as standing in contrast to all the other options the author had. Elements in the topic may or may not be contrastive, depending on the communicative purpose. Additionally, TFA incorporates the notion of *communicative dynamism* from Firbas’s Functional Sentence Perspective.

As can be seen, TFA is a fairly complex framework. The annotation used in the PDT does not capture the full complexity of TFA. Instead, it offers an abridged version in which each tree node in the tectogrammatical layer is annotated with one of three values:

- “t” is assigned to contextually bound, non-contrastive nodes.
- “c” is assigned to contextually bound, contrastive nodes.
- “f” is assigned to contextually unbound nodes (contrastive and non-contrastive).

As is apparent from the choice of letters, there is an unspoken assumption behind this scheme that contextually bound elements constitute the topic and contextually unbound elements constitute the focus. The topic is further subdivided into contrastive and non-contrastive sections, while the focus is not as it is understood to be contrastive by default. Additionally, the relative communicative dynamism of nodes is indicated in the corpus by their left-to-right order in the tectogrammatical tree.

The information structure annotation has been added to the PDT relatively recently and only covers a subset of the corpus. Nevertheless, it has been exploited for linguistic research already and some valuable findings have begun to emerge.

One obvious question to ask is whether contextually bound elements actually do refer to something previously mentioned in the text. Theoretically, a great deal of them should, as that is part of their definition. The PDT contains annotation for coreference (also at the tectogrammatical layer) and these two annotations can be cross-checked to verify the hypothesis. Unfortunately, the answer is inconclusive yet because the coreference annotation in the PDT is known to be incomplete in the sense that it fails to capture the full gamut of coreference phenomena of authentic texts. Still, whatever evidence there is suggests that the hypothesis may indeed be true: of all anaphoric links, 98.6% lead from contextually bound nodes (Kučová at al. 2005, p. 7). If the hypothesis is eventually demonstrated to be true, it will mean that the two annotation schemes (coreference and information structure) employed in the PDT are consistent with each other, and therefore probably correct as theories of how an aspect of (the Czech) language works. At this stage, it appears that the current mild inconsistency is caused by the inadequacy of the coreference annotation scheme, which is already known. That by itself is a valuable outcome of

the corpus annotation endeavour as it provides concrete evidence that the descriptive framework needs improvement – and that is another example of how corpora can be used to validate and improve existing linguistic theories.

The two discourse phenomena of information structure and coreference are interconnected in many important ways. For example, it has been suggested (Sgall et al. 2003, p. 7) that when resolving anaphora, all potential targets have a certain *salience*, and that salience is increased if the potential target is in the topic of a sentence rather than in the focus. It appears that the PDT might be the right place to verify this theory, and work on that is in progress (see Sgall et al. 2003).

It is probably obvious from the account so far that research into the interplay between information structure and coreference has only just started. That it is possible for researchers to investigate these questions now is thanks to the fact that a corpus has been made available with the appropriate annotations.

## **6. Conclusions**

Some linguistic phenomena are very difficult to investigate in a corpus without annotation – for example, it would have been extremely tedious for Dorothy Kenny to investigate the relationship between coreference and information structure in her raw corpus. Once a corpus is annotated, some important observations can be made by simply allowing the findings to emerge from the relevant statistics. This is true of all levels of linguistic description, but particularly so of the discourse level as it is at the furthest remove from the actual raw text.

The historical development of corpus linguistics has so far been such that the discipline progressively moves upwards from one level of linguistic description to the next. The discourse level has only been reached recently. Although corpus-based research into discourse is relatively new, the tentative results are very promising and strongly suggest that further investment into the discipline will yield important contributions towards a more complete understanding of how people make meaning through language – both in the form of language-specific observations and in the form of the validation (or improvement, or even rejection) of pre-existing theoretical descriptive frameworks.

At this early stage of its development, corpus-based discourse analysis is most firmly rooted in the analysis of spoken language, and in types of analysis that do not go very far beyond the sentence boundary. But discourse features are present in

written texts too, even if not so obviously, and some discourse features reach very far beyond individual sentences: for example, the information structures of individual sentences come together to form an information structure for the whole text. Such long-distance discourse patterns have not been investigated by corpus linguistics yet, but that is only a matter of time. In the future, corpus linguistics is likely to expand into all areas of discourse analysis.

## References

- Firbas, J. (1992) *Functional Sentence Perspective in Written and Spoken Communication* Cambridge: Cambridge University Press
- Hajičová E.; J. Panevová; P. Sgall (2003) *Úvod do teoretické a počítačové lingvistiky, I. svazek - Teoretická lingvistika [Introduction to Theoretical and Computational Linguistics, Volume I - Theoretical Linguistics]* Prague: Karolinum
- Hajičová E.; P. Sgall (2004) 'Degrees of Contrast and the Topic-Focus Articulation' <<http://ufal.mff.cuni.cz/pdt2.0/publikaci/HajicovaSgalI2004.pdf>> accessed 1 December 2007
- Halliday, M.A.K (1994) *An Introduction to Functional Grammar, 2nd edition* London: Edward Arnold
- Kenny, D. (1998) 'Theme and Rheme in Irish and English: a corpus-based study' in Cronin, M. (ed.) *Working papers in language and society* Dublin: School of Applied Language and Intercultural Studies, Dublin City University
- Kučová L.; E. Hajičová; K. Veselá; J. Havelka (2005) 'Topic-focus articulation and anaphoric relations: a corpus based probe' in *Prague Bulletin of Mathematical Linguistics*, 84 Prague: Faculty of Mathematics and Physics, Charles University. Also available at <<http://ufal.mff.cuni.cz/pdt2.0/publikaci/KucovaHajicovaVeselaHavelkaPML2005.pdf>> accessed 1 December 2007
- Sampson, G. (1995) *English for the Computer: The Susanne Corpus and Annotation Scheme* Oxford: Clarendon Press
- Sgall, P.; E. Hajičová; E. Buráňová (2003) 'Topic-Focus Articulation and degrees of salience in the Prague Dependency Treebank' in A. Carnie; H. Harley; M. Willie (eds.) *Formal Approaches to Function in Grammar* Amsterdam and Philadelphia: John Benjamins. Also available at <<http://ufal.mff.cuni.cz/pdt2.0/publikaci/HajicovaSgalIBuranova2003.pdf>> accessed 1 December 2007

Widdowson H.G. (2007) *Discourse Analysis* Oxford: Oxford University Press

*Corpora*

*Susanne* <<http://www.grsampson.net/RSue.html>> accessed 2 December 2007

*Childes* <<http://childes.psy.cmu.edu/>> accessed 2 December 2007

*Prague Dependency Treebank version 2.0* <<http://ufal.mff.cuni.cz/pdt2.0/>> accessed  
1 December 2007