# Towards a Corpus-lexicographical Discourse Analysis

Emma Franklin

Lancaster University, Bailrigg, Lancaster, LA1 4YW

**Abstract.** This working paper presents the progress made thus far in the development of a corpus-lexicographical approach to discourse analysis, more specifically the application of Hanks' [5, 6] Corpus Pattern Analysis (CPA) procedure to a (critical) discourse analysis task. The theoretical basis of CPA is explained, followed by some practical applications of CPA, namely lexicography and the proposed method of discourse analysis. Examples are taken from an ongoing investigation into the use of 'killing' verbs in contemporary British English, which draws upon two corpora: the British National Corpus (BNC) and the animal-themed 'People', 'Products', 'Pests' and 'Pets' (PPPP) corpus [8]. Preliminary findings suggest that a CPA-assisted, or corpus-lexicographical, discourse analysis is one with a strong theoretical basis, whose transparency and systematicity empowers the analyst to make precise and persuasive arguments.

**Keywords:** Discourse Analysis, Corpus Pattern Analysis, Lexicography.

## 1 Introduction

Current methods of discourse analysis are numerous and wide-ranging, to the point that such terms as "discourse analysis" and even "critical discourse analysis" are almost meaninglessly vague. The word "discourse" is, itself, polysemous, and the aim of this paper is not to attempt to untangle its nuances. Rather, this work seeks to carve out a new, potential route to understanding meaning in discourse using corpus-lexicographical methods, and some of the progress made thus far in this approach is presented here.

In over-simplified terms, for the purposes of this brief discussion, "discourse" is understood to refer in its non-countable form to "language in use", and in its countable form – or "big D" form [3] – to a "conventional practice". Critical Discourse Analysis (CDA) is defined most simplistically as "discourse analysis 'with attitude'" [11, p. 96]. More specifically, it seeks "to uncover and de-mystify certain social processes in this and other societies, to make mechanisms of manipulation, discrimination, demagogy and propaganda explicit and transparent" [12, p. xiv].

Though discourse is arguably language above the level of sentence or clause, it is constituted by much smaller units of language which need to be analysed as such. Corpus-assisted discourse analyses already take this route, traditionally via the use of statistically generated word lists, collocates, keywords, and so on [2]. The approach proposed here does not make use of most of these methods, but does rely on corpus data and uses corpus analysis software to generate concordance lines for manual inspection. It employs Hanks' [5, 6] Corpus Pattern Analysis procedure, the output of which is

considered in light of existing literature as well as historical and political context. The result is an empirical, semantically motivated, critical analysis of argument structure across text types. It is lexicographical in that it entails the creation of a corpus-based lexicographical entry for a given word, as per the *Pattern Dictionary of English Verbs* (PDEV), introduced in Section 2. It does not currently make use of automated natural language processing, e.g. semantic parsing, but instead relies on manual analysis for accurate classification of arguments and delimitation of word senses.

The rest of the paper is organised as follows. Section 2 outlines the theoretical background and main features of Corpus Pattern Analysis (CPA), followed by some examples of practical applications of CPA, namely lexicography and corpus-lexicographical discourse analysis, including a short case study. Section 4 concludes the paper with a very brief summary of the potential rewards and challenges of taking this approach.

## 2 Corpus Pattern Analysis and the PDEV

Corpus Pattern Analysis (CPA), developed by lexicographer Patrick Hanks, seeks "[to elucidate] the relationship between syntagmatic patterns and activated meanings" [5, p. 92]. Following in the Neo-Firthian tradition, CPA examines the behaviour of words in their contexts, and in doing so establishes the linguistic patterns with which word senses are associated. Words, Hanks argues, do not have meaning but "meaning potential"; their meanings are only activated by the lexical patterns in which they exist [5] and, like Sinclair, Hanks finds meaning to be inextricably linked to form (cf. [10]). So far, CPA has mostly been employed in lexicography, namely the *Pattern Dictionary of English Verbs*[1], under the *Disambiguation of Verbs by Collocation* (DVC) project[2].

CPA is underpinned by Hanks' [6] Theory of Norms and Exploitations (TNE), which centres on the phenomenon of prototypical language use (norms) and exploitations of these norms. CPA takes a similar approach to that of the COBUILD [9] and Hector [1] projects, and bears some similarities to Construction Grammar [4]. However, CPA is more concerned with lexical semantics, and it relies wholly on corpus evidence of usage. A *pattern*, in the CPA sense, "consists of a valency structure … together with sets of preferred collocations" [6, p. 92]. Patterns can be *norms* (patterns of normal, conventional, everyday usage) or *exploitations* (creative patterns of language use), though the distinction between the two is not an absolute one [6, p. 4].

Following Pustejovsky [7], CPA employs *semantic types*, which are logical constructs for groups of lexical items, arranged in a hierarchical semantic ontology. For example, the verb *sip* selects as its direct object lexical items such as *beer*, *water*, *whiskey*, and *tea*, which form a lexical set represented in the *CPA Ontology*[3] by the semantic type of [[Beverage]]. A [[Beverage]] is a [[Liquid]] is a [[Fluid]] is [[Stuff]] is an [[Inanimate]] is a [[Physical Object]], and so on. The CPA Ontology is unique, in that it was not devised *a priori*, but instead was progressively built and altered during the

[1] http://pdev.org.uk
[2] http://gtr.rcuk.ac.uk/projects?ref=AH/J005940/1
[3] http://pdev.org.uk/#onto

course of the project, and can be considered to be data-driven and specific to the corpus upon which it is based (the British National Corpus (BNC), predominantly).

The semantic types from the CPA Ontology occupy argument slots, for example, the subject, object and prepositional object slots. CPA patterns are anchored to *implicatures*, which form an integral part of a word's "syntagmatic profile" [6] and which describe the entailment of a particular pattern. For example, the most common pattern associated with the verb *drink* is listed in PDEV as (1), with the implicature, (2).

[[Human]] drink [[Beverage]] ({up | down}) (1)

[[Human]] takes [[Beverage]] into the mouth and swallows it (2)

Words in double square brackets are semantic types. The round brackets in (1) denote optionality, i.e. in this instance, an adverbial is not always present. Curly brackets denote specific lexical items; in this case, *up* and *down* cannot be substituted.

Finally, it should be noted that CPA is concerned with conventionality; it does not classify what is *possible* in language, but what is *typical*. CPA patterns, like semantic types, represent central, canonical forms of language as opposed to all potential ones.

## 3 Putting CPA into Practice

### 3.1 Doing Lexicography with CPA

The standard CPA procedure is described in detail elsewhere [5, 6]. To summarise:

- The analyst generates a concordance for a node word and takes a random sample of concordance lines, starting with around 250. In the interests of producing generalisable results, a large, general-language corpus is used as a source of data.
- Lines are manually grouped together based on their shared syntagmatic properties – their valency, arguments, presence or absence of adverbials, etc. This involves identifying norms (prototypical phraseology) and from there deciding which instances are likely to be exploitations. Establishing such patterns "calls for a great deal of lexicographic art" [5, p. 88].
- The analyst sorts these grouped lines into patterns by tagging each line with a pattern number, and then writing up the patterns and their implicatures into a kind of dictionary entry (see Fig. 1).

Using CPA for lexicography results in an empirically well-founded dictionary entry which gives the proportions of different word senses in the data. In other words, meaning becomes somewhat measurable. Lexicography that more accurately represents natural language use is valuable not only for language learners and teachers, but also for computational linguists interested in semantic probabilities for the purposes of word-sense disambiguation. Measuring the presence of word senses in "general" language also makes it possible to compare meanings across texts.

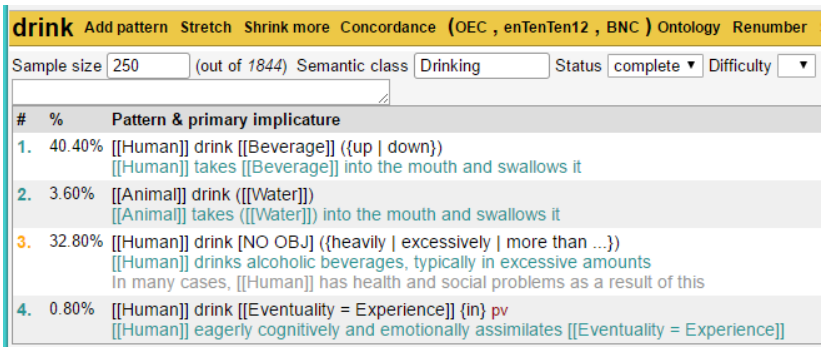


**Fig. 1.** Non-public-facing PDEV entry for the verb *drink*.

### 3.2 Doing Discourse Analysis with CPA

CPA for lexicographical purposes involves the use of large, general, reference corpora, such as the BNC. As discourse analysts tend to be interested in one particular type of discourse, or how discourses differ from one another, a corpus-lexicographical discourse analysis will also involve carrying out CPA on a specialised corpus or subcorpus. For highly specialised or technical language, a new ontology of semantic types may have to be created from scratch. In most cases of contemporary British English investigations, however, the PDEV's CPA Ontology will act as a useful starting point.

By way of example, my ongoing doctoral research is a corpus-assisted investigation into 'killing' phraseology in contemporary British English, with a particular focus on human-animal relations and how 'killing' events are represented across discourses. Killing is a process involving multiple participants, e.g. agents and patients, or 'killers' and 'killees', making verbs an ideal place to start; predicates act as the pivot of a clause, and so to analyse a verb is to uncover the arguments it governs. As CPA is systematic, empirical, and particularly well-suited to verbs, it forms the basis of the analysis.
The specialist corpus used in this project is the 'People', 'Products', 'Pests' and 'Pets' (PPPP) corpus [8]. It comprises almost 9 million words of animal-related discourse in contemporary British English from a range of text types and genres; see Table 1 for the composition details. The BNC is used as a reference corpus, and PDEV entries are referred to where available. Corpus software AntConc[4] is used for generating concordances, and Microsoft Excel is used for sampling, tagging, sorting and analysing.

The procedure for CPA-assisted discourse analysis, in this project, is as follows:

- Consult the PDEV to see whether the verb in question already has an entry. If not, take a 250-line random sample of the verb's concordances from a POS-tagged version of the BNC and carry out CPA using the CPA Ontology.
- Take a 250-line random sample of the verb's concordances from the POS-tagged version of the PPPP corpus and carry out CPA, using the PDEV/BNC patterns as

[4] http://www.laurenceanthony.net/software.html

a loose guide. The CPA Ontology is a basis from which to start creating a new ontology tailored to the specialised corpus over time; this is an iterative process.

- Compare occurrence and distribution of patterns across the two samples, and observe differences in semantic types. Discuss these in context of the literature.

**Table 1.** PPPP Corpus Composition, from [8].

| Subcorpus | No. of files | No. of types | No. of tokens |
|---|---|---|---|
| Broadcasts | 83 | 19835 | 614378 |
| Campaign literature | 470 | 16488 | 306680 |
| Legislation | 843 | 10201 | 627127 |
| Food websites | 258 | 7503 | 87118 |
| Journals | 1609 | 93567 | 5698531 |
| News | 1023 | 28777 | 466340 |
| MO Project contributions | 103 | 9931 | 174938 |
| Focus groups | 19 | 8277 | 229059 |
| Interviews with text producers | 17 | 8068 | 157664 |
| Interviews with dog keepers | 19 | 8698 | 309719 |
| Total | 4444 | 211345 | 8671554 |

## Case study: *destroy*

Given that it refers in some contexts to killing, *destroy* was selected as a candidate for analysis. The above steps were carried out and the same patterns were found in both the PPPP and the BNC samples, though in different proportions. Pattern 1, which refers to the attacking or damaging of a physical object, is equally prominent in both samples. Patterns 2 and 3, which refer to abstract senses of destruction (e.g. of confidence, and a human opponent, respectively), are less prominent in the PPPP corpus sample. The proportion of Pattern 4, which refers to the killing of animals (and fetuses) by humans, is (as expected) far higher in the PPPP sample. See Fig. 2 for the pattern details.

| # | BNC | PPPP | Pattern & primary implicature |
|---|---|---|---|
| 1 | 60.40% | 59.49% | [[Human \| Animal \| Institution \| Event \| Artifact]] destroy [[Physical_Object]]<br>[[Human \| Animal \| Institution \| Event \| Artifact]] damages or attacks [[Physical_Object]] until it is completely ruined |
| 2 | 28.00% | 5.06% | [[Anything]] destroy [[Property \| Abstract_Entity \| State_of_Affairs]]<br>[[Anything]] causes [[Property \| Abstract_Entity \| State_of_Affairs]] to no longer exist |
| 3 | 5.60% | 1.27% | [[Human 1]] destroy [[Human 2 \| Human_Group]]<br>[[Human 1]] utterly defeats [[Human 2 \| Human_Group]] |
| 4 | 3.20% | 34.18% | [[Human]] destroy [[Animal \| Animal_Group \| Fetus]]<br>[[Human]] kills unwanted [[Animal \| Animal_Group \| Fetus]] |

**Fig. 2.** Patterns and implicatures for the verb *destroy*, as found in the BNC and PPPP samples

CPA makes it possible to distinguish not only inter-pattern differences (the pattern boundaries) but also intra-pattern variation, including anomalies. While pattern distribution across corpora is useful, the real value to discourse analysis is found within the boundaries of the patterns themselves. For example, *destroy* in Pattern 1 takes as its object a Physical Object. These are typically physical objects of the inanimate variety, particularly artifacts and buildings, which is intuitive given the etymology of *destroy* (from the Latin *destruere*, lit. "unbuild"). In the BNC these tend to be houses and vehicles, and in the PPPP corpus these are more often nests, setts and other animal homes.

Less predictable and less straightforward to deal with are those examples which feature unusual arguments. Take, for example, lines (3-6), found in the PPPP sample.

Did a meteorite really *destroy* the dinosaurs? (3)

[…] otters have *destroyed* entire populations of large fish in some fisheries […] (4)

[…] everything will be *destroyed*, the animals, the plants, the water, the land. (5)

Scottish Ministers may (a) cause to be *destroyed* any semen, egg or embryo […] (6)

Having carried out CPA, we can say that although these examples make sense and are *possible*, they are not particularly normal or *typical*. Given that all four involve killing, and not merely damage, we might consider them instances of Pattern 4. However, Pattern 4 refers to the killing of (unwanted) animals by humans, usually in an official, procedural context. Hence, (3) and (4) must instead belong to Pattern 1; they certainly do not fit with Patterns 2 and 3. The same goes for (5) and (6); although they involve killing, they are not describing the sort of event typically construed by Pattern 4. Their objects include animals and fetuses, but they also involve inanimate objects. Verb senses do not change mid-argument, unless in a creative exploitation such as wordplay [6, p. 72]; therefore, the same sense of the verb must apply to all entities in this co-hyponymy. If 'water' and 'eggs' cannot be killed, then we know that these examples are referring instead to destructive *damage*, i.e. belong to Pattern 1. Using CPA, it is therefore possible to say that while animals are sometimes *destroyed* in the same sense as inanimate objects such as houses and cars, humans are not. This is an assertion now provable with evidence.

In terms of my research, this finding contributes to answering the research question, "In what ways are animals conceptualised as persons, and in what ways are they conceptualised as things?", which can only be answered in full once evidence has been gathered from a wide range of 'killing' verbs and their patterns. Nevertheless, this example is one small step towards demonstrating the inherent anthropocentrism of (English) language, and the persistent, widespread and insidious habit of likening nonhuman beings to inanimate things. In light of literature found in human-animal studies and critical animal studies, this might feasibly be interpreted as an attempt by the speaker to justify the routine exploitation of animals by humans, or at the very least as a betrayal of their view – subconscious or otherwise – that some animals are more similar to insentient objects than they are to humans, and as such deserve their subordinate status.

## 4 Conclusions

The very brief example given in Section 3.2 is not a full analysis, owing to space limitations, but it points towards a route along which CPA might be used as an empirical basis for (critical) discourse analysis. The systematicity of the CPA procedure ensures that conclusions drawn from the data are reached methodically and with measurable evidence. I might assert, for instance, that humans objectify other animals with their language, and I might – as has been done by many in human-animal studies – be able to provide several examples of this type of oppressive language. However, such assertions are made less convincing by not having accounted for the whole picture (the 'whole picture' being, admittedly, a sample of the whole picture). By starting with a word and using CPA to map the meanings of that word across large samples of text, we can demonstrate with more accuracy where, when and in what ways its norms are being used and exploited, often subtly and even unknowingly, to further a particular ideology.

There are some challenges to using this approach. It currently involves a lot of manual tagging, and at its most effective CPA requires a corpus-specific ontology of types, which takes time and dedication. However, once the groundwork has been laid, the discourse analyst has an empirically well-founded and robust basis from which to explore meaning. The classification of concordance lines in terms of their arguments rather than surface-level representations has advantages in terms of generalisability, and the use of CPA does not preclude – rather bolsters – other forms of textual analysis.

## References


1. Atkins, S.: Tools for Computer-aided Corpus Lexicography: the Hector Project. Acta Linguistica Hungarica, 41, 5-72 (1993).
2. Baker, P.: Using Corpora in Discourse Analysis. Continuum, London (2006).
3. Gee, J.: Social Linguistics and Literacies: Ideology in Discourses. 5th edn. Routledge, London (2015).
4. Goldberg, A. E.: Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press, Chicago (1995).
5. Hanks, P.: Corpus Pattern Analysis. In: Proceedings of the 11th EURALEX International Congress, pp. 87-97, Lorient, France (2004).
6. Hanks, P.: Lexical Analysis: Norms and Exploitations. MIT Press, Cambridge, MA (2013).
7. Pustejovsky, J.: The Generative Lexicon. Computational Linguistics 17(4), 409-441 (1991).
8. Sealey, A., Pak, C.: First catch your corpus: methodological challenges in constructing a thematic corpus. Corpora (forthcoming).
9. Sinclair, J.: Looking Up: An Account of the COBUILD Project in Lexical Computing. Collins, London (1987).
10. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991).
11. van Dijk, T.: Multidisciplinary CDA: a plea for diversity. In: Wodak, R., Meyer, M. (eds.) Methods of Critical Discourse Analysis. Sage, London (2001).
12. Wodak, R.: Language, Power and Ideology: Studies in Political Discourse. Critical Theory, Vol. 7. John Benjamins, Amsterdam (1989).