# On the use of automatic tools for large scale semantic analyses of causal connectives

**Liesbeth Degand**
Université catholique de Louvain
Institute of Linguistics
degand@lige.ucl.ac.be

**Wilbert Spooren**
Vrije Universiteit, Amsterdam
Language & Communication
w.spooren@let.vu.nl

**Yves Bestgen**
Université catholique de Louvain
Faculty of Psychology
Yves.bestgen@psp.ucl.ac.be

## Abstract

In this paper we describe the (annotation) tools underlying two automatic techniques to analyse the meaning and use of backward causal connectives in large Dutch newspaper corpora. With the help of these techniques, Latent Semantic Analysis and Thematic Text Analysis, the contexts of more than 14,000 connectives were studied. We will focus here on the methods of analysis and on the fairly straightforward (annotation) tools needed to perform the semantic analyses, i.e. POS-tagging, lemmatisation and a thesaurus-like thematic dictionary.

## 1 Introduction

In ongoing work, we explore the possibility to make use of large corpora to test hypotheses concerning linguistic factors determining the meaning and use of connectives. Of course, corpus-based approaches of connectives are not new, but classically they consist of either fully analysed but relatively small corpora, or of large corpora of which a random set is analysed. The reason for this quantitative restriction is clear: The data-analyses are completely hand-based. While these empirical studies are useful from a qualitative point of view, they all suffer from the same quantitative drawback, namely the relatively small number of data (rarely more than 100 occurrences are analysed, mostly only 50). In addition, most of these analyses are still too analyst-dependent, making generalizations and replications difficult. Changing this situation includes handling exhaustively large corpora (with hundreds and even thousands of occurrences of the same linguistic phenomenon) and implementing the analytic procedures to make them analyst-independent. In this paper, we test such a methodology for which we used a number of linguistic hypotheses found in the literature on the semantics of causal connectives and tried to replicate the results. The linguistic material we worked on are four Dutch backward causal connectives: *aangezien* ('since'), *doordat* ('because of the fact that'), *omdat* ('because') and, *want* ('because'). This choice was motivated by the fact that there has already been quite some linguistic work on this topic, mainly empirically based (Degand, 2001; Degand and Pander Maat, 2003; Pit, 2003).[1] We have shown elsewhere how linguistic hypotheses concerning the scaling of these connectives in terms of subjectivity and their thematic behaviour could be supported (Bestgen et al., 2003). Since these first results are very encouraging, we would like to focus here on the methods of automatic analysis – Latent Semantic Analysis and Thematic Text Analysis - and on the fairly straightforward (annotation) tools needed to perform the semantic analyses, i.e. POS-tagging, lemmatisation and a thesaurus-like thematic dictionary. We illustrate how the combination of the two techniques of automatic analysis permit to gain deeper insight into the semantic constraints on the use of the connectives studied. Doing so, we test a number of new hypotheses concerning the perspectivizing and polyphonic nature of connectives that remain unconfirmed in the linguistic literature. We also discuss the robustness of the techniques and their reusability in other contexts and other languages.

[1] For lack of space we will not present the linguistic analyses here but will consider them as given.

## 2 Techniques and Tools

The techniques used have to fulfil two tasks: they are needed to extract the relevant linguistic material from the corpus, that is to say the four connectives with their context of use; and they are used to analyse the retrieved elements in order to test a number of linguistic hypotheses concerning the meaning and use of these connectives. Our main objective is to show that with the use of these techniques only fairly straightforward annotation tools are needed to perform quite profound semantic analyses on massive quantitative data.

### 2.1 POS-Tagging and the identification of the causal segments

The extraction of the relevant linguistic material was fulfilled by automatic syntactic analysis techniques. As a basis for our analyses we worked with the first six months of a Dutch newspaper corpus of more than 30 million words[2]. This material was POS-tagged using MBT (Memory Based Tagger) (Daelemans et al.,1996). We then discarded the items with few content words: sports results, television programs, crosswords and puzzles, stock exchange reports, service information from the newspaper editor, etc. We also ‘cleaned’ the corpus material of irregularities caused by the incompatibility between the source file and the tagging program (mostly nonsense words generated by the program). This eventually led to a data set of approximately 16,500,000 words.

The POS-tagging permitted to segment the corpus in sentences and to label the words grammatically. Second, POS-tagging allowed us to locate and extract the connectives from the sentences in which they occurred. Concretely, we extracted all sentence-length segments on the basis of the tag <UT> (‘utterance’). We then did a search on the four connectives tagged as <conj> by the parser.

Table 1 displays the frequencies of the retrieved connectives. These figures do not include a number of sentences that were eliminated because they were potentially problematic for the analysis. This was for instance the case for sentences containing more than one connective out of our list of four.[3]

| **Connective** | **Raw frequency** | **Relative frequency (per million words)** |
|---|---|---|
| aangezien | 248 | 30 |
| Doordat | 826 | 101 |
| Omdat | 7689 | 938 |
| Want | 5621 | 686 |

Table 1: Frequencies of the causal connectives in the data set

The extracted sentences were then analysed in terms of a series of heuristics to identify the CAUSE (P) and CONSEQUENCE segments (Q)[4]. From a syntactic point of view, the connectives *doordat, omdat* and *aangezien* can occur in two basic types of causal constructions: medial (Q CONNECTIVE P), see example (1), and preposed ones (CONNECTIVE P, Q), see example (2). The connective *want* only appears in medial constructions.

(1) Een gezamenlijk beleid is nodig **omdat** in het najaar in het Japanse Kyoto wereldwijd wordt onderhandeld over het klimaat.
‘A common policy is necessary because worldwide negotiations will take place in the autumn in the Japanese city of Kyoto.’

(2) Iedere strenge winter heeft gevolgen voor de kerkorgels', zegt dr. A.J. Gierveld van de Gereformeerde Organisten-vereniging. **Doordat** het hout krimpt, kunnen er kieren ontstaan waardoor lucht ontsnapt.
‘Every hard winter has consequences for the church organs”, Dr. A.J. Gierveld of the Reformed Organists Union says. Because the wood shrinks, crocks may show, through which air escapes.’

The heuristics to identify the CAUSE (P) and CONSEQUENCE (Q) segments were primarily based on

[2] We used the year 1997 of "De Volkskrant" a Dutch national daily newspaper. The corpus is distributed on CD-rom.

[2] These cases were eliminated in order to be sure of the exact influence of the connective and about the exact contribution of the context.

[4] The connectives under investigation are all so-called backward causal connectives, i.e. they express an underlying causal relation of the type CONSEQUENCE – CAUSE, in which the connective introduces the CAUSE segment.

a) the position of the connective in the sentence (number and type of words preceding the connective),

b) the number, position and order of finite verbs in the segment,

c) the presence or absence of punctuation markers, especially commas.

For example, a sentence beginning with the connective *omdat* can either be preposed (P-Q) (example 3), or medial (Q-P), if Q and P are given in different sentences (example 4).

(3) **Omdat** de verdachte niet eerder was veroordeeld, bleef de gevangenisstraf geheel voorwaardelijk.
'Because the suspect had not been convicted before, the sentence was entirely probational.'

(4) Maar er zijn meer programma's die de moeite waard zijn en die toch niet worden bekeken. **Omdat** ze onvindbaar zijn tussen de ramsj.
'But there are more [TV] programmes that are worth watching and still are not being watched. Because they are hard to trace among the rubbish.'

To extract these segments correctly, a number of rules enter into play. For example,

a) If CONN = omdat, doordat or aangezien; and

b) If CONN in initial position, look for first finite verb [vf], if vf appears in segment <…vf, vf …> or <… vf vf …>, then cut before second vf, and segment containing CONN is P, the other one is Q.

c) If CONN in initial position and there is only one vf, then segment containing CONN is P, and previous sentence is Q.

Other rules are used to determine whether the CONN is in initial position or not. In addition to examples (2-3), example (5) also illustrates a case of initial connective, even though a word precedes the connective.

(5) En **omdat** in Nederland de voertaal nog steeds het Nederlands is, worden de meeste schoolvakken ook in die taal gedoceerd.
'And because Dutch is still the main language in the Netherlands, most subjects are taught in that language.'

This resulted in 21 heuristic rules, the adequacy of which was hand-checked on large samples of the data. In the end, 1.4% of the data were lost because one of the segments was missing or because none of the procedures could work out the identification of P and Q. Ultimately we were able to identify the causal segments for 14181 sentences. Four syntactic environments can be distinguished, involving a preposed construction <Conn P Q.> as in examples (2, 3, 5) above, and three types of medial constructions:

a) <Q conn P.> corresponds to a construction in which Q and P are linked by a connective within the same sentence (example 1);

b) <Q. Conn P.> corresponds to constructions in which the previous sentence functions as Q (examples 4); and

c) <Prev. Q conn P.> corresponds to constructions for which the Q-segment is anaphoric with the preceding sentence, thus requiring this previous sentence for the semantic interpretation, as in example (6), in which the Q “dat komt” (litt. ‘that comes’) picks up the semantic information from the previous sentence and links it to the P-segment introduced by the connective.

(6) De Europese economie raakt hopeloos achterop bij de Amerikaanse en Japanse. **Dat komt doordat** Europa niet meedoet op nieuwe groeimarkten.
'The European economy is falling hopelessly behind the American and Japanese economy. This is because

```
        Europe is not participating
        in new growth markets.’
```

Actually, only 7.1% of the sentences investigated belong to the preposed construction type. However, important divergences exist between the connectives: *want* is never used in preposed position, *omdat* in 10.41% of the cases, and *doordat* in 14.32% of the cases, a figure which rises to 43.5% of the cases for *aangezien*. It is interesting to point out that this is in total agreement with previous small-scale corpus research on this matter.

## 2.2 Lemmatisation and the construction of the LSA semantic space

The first automatic technique that will be presented is Latent Semantic Analysis (LSA), a mathematical technique for extracting a very large “semantic space” from large text corpora on the basis of the statistical analysis of the set of co-occurrences in a text corpus. Landauer et al. (1998) stress that this technique can be viewed from two sides. At a theoretical level, it is meant to be used to develop simulations of the cognitive processes running during language comprehension, including, for instance, a computational model of metaphor treatment (Kintsch, 2000 ; Lemaire et al., 2001), but also to analyse the coherence of texts (Foltz et al., 1998 ; Piérard et al., 2004). At a more applied level, it is a technique which enables to infer and to represent the meaning of words on the basis of their actual use in text so that the similarity of the meaning of words, sentences or paragraphs can be estimated (Bestgen, 2002; Choi et al., 2001). It is this latter aspect which draws our attention here.

The point of departure of the analysis is a lexical table (Lebart and Salem, 1992) containing the frequencies of every word in each of the documents included in the text material, a document being a text, a paragraph, or a sentence. To derive semantic relations between words from the lexical table the analysis of mere co-occurrences will not do, the major problem being that even in a large corpus most words are relatively rare. Consequently the co-occurrences of words are even rarer. This fact makes such co-occurrences very sensitive to arbitrary variations (Burgess et al., 1998 ; Kintsch, 2001). LSA resolves this problem by replacing the original frequency table by an approximation producing a kind of smoothening effect on the associations. To this end, the frequency table undergoes a singular value decomposition and it is then recomposed on the basis of only a fraction of the information it contains. Thus, the thousands of words from the documents have been substituted by linear combinations or ‘semantic dimensions’ with respect to which the original words can be situated again. Contrary to a classical factor analysis the extracted dimensions are very numerous and non-interpretable.

All original words and segments can then be placed into this semantic space. The meaning of each word is represented by a vector, thus indicating the exact location of the word in this multidimensional semantic space. To calculate the semantic proximity between two words, the cosine between the two vectors that represent them is calculated. The more two words are semantically similar, the more their vectors point in the same direction, and consequently, the closer their cosine will be to 1 (coinciding vectors). A cosine of 0 shows an absence of similarity, since the corresponding vectors point in orthogonal directions. It is also possible to calculate the similarity between ‘higher order’ elements, i.e. between sentences, paragraphs, and entire documents, or combinations of those, even if this higher order element isn’t by itself an analysed element. The vector in question corresponds to the centroid of the words composing the segment under investigation. The centroid results from the weighted sum of the vectors of these words (Deerwester et al., 1990). This makes it possible to calculate the semantic proximity between any two sentences, viz. whether present in the original corpus or not, whether the original corpus had been segmented in sentence length documents or not.

To perform the LSA analyses, we used the Dutch newspaper corpus to build the semantic space. To this end, the data set, which had been lemmatised with MBLEM (Memory Based Lemmatiser) (Van den Bosch & Daelemans, 1999), was cut into article-length segments. Elimination of all digits, special characters, punctuation marks, and of a list of 222 stopwords (words occurring in “any” context, like determiners, auxiliaries, conjunctions, …), brought the total number of words back to approximately 6.5 million. For the input lexical table, the documents were articles of minimally 24 words and maximally 523 words, i.e. all articles minus the 10% shortest and minus the 10%

longest ones. As to the words, we kept all those that occurred at least ten times in the data set. Overall this resulted in a matrix of 36630 terms in 28640 documents. To build the semantic space proper, the singular value decomposition was realized with the program SVDPACKC (Berry, 1992; Berry et al., 1993), and the 300 first singular vectors were retained. In the present research we will use this technique to evaluate the semantic proximity between P& Q, and between the causal segments and the prior or subsequent sentences.

### 2.3 Dictionaries and lexical categorisation

The second technique used to test the linguistic hypotheses is alternatively called ‘word count strategy’ (Pennebaker et al., 2003), automatic identification of linguistic features (Biber, 1988) or thematic text analysis (Popping, 2000; Stone, 1997), the aim of which is to determine whether some categories of words (e.g., words of opinion, fact, attitude, etc.) or some grammatical categories (e.g. personal pronouns) occur more often in a given type of text segment. The first step in this kind of analysis is to build a dictionary that contains the categories to be investigated and the corresponding (lemmatised) lexical entries that signal their occurrence. The categories may correspond to grammatical classes, but also to thematic word grouping. The following step consists in searching all the text segments containing these lexical entries in order to account for the frequency of each category in each text segment. These data are put into a matrix that has one row for each text segment and one column for each category, each cell containing the frequency of the respective category in the respective text segment. Finally, this matrix is analysed to determine whether some categories occur more often in a given type of text segment. To illustrate this technique, let us assume that we want to test the hypothesis that (nominative) personal pronouns occur more frequent in text segments connected by *want* than by the other backward causal connectives. In the first step the "Personal-Pronoun" dictionary is built, containing the corresponding lexical entries: *ik, jij, je, hij, zij, ze, wij, we, jullie, u.* All the text segments containing these lexical entries are then searched in order to account for the frequency of the concept "Personal-Pronoun" in each text segment. These data are put into a matrix which is analysed to determine whether the concept "Personal-Pronoun" occurs more often with *want*-segments than with the other causal segments.

The two main difficulties we are confronted with when using this technique in the present studies are (i) the reduced size of the analysed text segments (one sentence or even less), and (ii) the difficulty, or even impossibility, to build an exhaustive list of words belonging to a category like fact, opinion, attitude, etc. With respect to the first difficulty, we believe that the reduced size of the segments will be compensated by the large number of segments of each type being analysed. The second difficulty is addressed below where we propose a number of ways to extend the category lists automatically.

## 3 Combining LSA and TTA: an application

### 3.1 Perspective shift

There are a number of claims in the literature that some connectives co-occur with perspective shifts between the causal segments, while others do not. Perspectivisation accounts for the fact that there are more sources of information than the speaker alone. In relation to our connectives, perspectivisation has been claimed to play a role in the meaning differences between *want* (introducing a perspective shift) and *omdat* (no perspective shift). However, the various corpus studies on this matter have not univocally confirmed this hypothesis (Degand, 2001; Oversteegen, 1997). We would like to explore this matter further by comparing the semantic tightness of the segments related by our connectives. This will be done by calculating the semantic proximity between Q and P for each of the connectives. Our hypotheses are as follows:

> Hypothesis 1: The cosine between Q and P related by monophonic connectives (*omdat*) should be higher than the cosine between Q and P related by polyphonic connectives (*want*).
>
> Hypothesis 2: The cosine between the prior sentence and the subsequent sentence should be higher for monophonic connectives than for polyphonic connectives.

| | Cos. Q & P | | Cos. Prior Subsequent | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| aangezien (N = 200) | 0.143 | 0.17 | 0.207 | 0.21 |
| doordat (N = 644) | 0.154 | 0.17 | 0.187 | 0.19 |
| omdat (N = 5691) | 0.137 | 0.17 | 0.182 | 0.20 |
| want (N = 3974) | 0.120 | 0.17 | 0.150 | 0.19 |

Table 2: Mean Cosine per connective between the causal segments, and between the prior and subsequent sentences

Table 2 displays the cosines resulting from the LSA-analysis. Two ANOVAs were performed. The first one had the connectives as independent variable and the semantic proximity between the causal segments as dependent variable. It shows that hypothesis 1 is borne out (F(3, 10505) = 11.36, p < 0.0001): the causal segments related by the (monophonic) connective *omdat* are semantically closer than the segments related by the (polyphonic) connective *want*. The results furthermore show that *doordat* and *aangezien* should be described in terms of monophonic connectives. The second ANOVA, with the connectives as independent variable and the semantic proximity between the prior and subsequent sentences as dependent variable, confirms hypothesis 2 (F(3, 10505) = 25.75, p < 0.0001): the monophonic connectives *aangezien*, *doordat* and *omdat* go along with topic continuity (or at least semantic proximity) between the prior and subsequent sentence to the causal construction, while this is less the case for the connective *want*.

To confirm that these results are indeed related to the issue of perspectivisation, this LSA-analysis was completed with a thematic text analysis to test for the presence vs. absence of perspective indicators. To this end we built a "Perspective" dictionary of perspective-indicating elements (Spooren, 1989) such as intensifiers, emphasisers, attitudinal nouns and adjuncts, etc. (Caenepeel, 1989). The dictionary was composed of two subcategories:

a) communication markers, like (non-ambiguous) verbs and adverbs of saying and thinking, e.g. *report, tell, confirm, require, according to,…*

b) markers of the speaker's attitude, like linguistic elements expressing an expectation or a denial of expectation, intensifiers and attitudinals, and evaluative words, e.g. *probably, must, horrible, fantastic, …*

To build the dictionary, we used a Dutch thesaurus (Brouwers, 1997) and extracted all (unambiguous) lemmas corresponding to one of the above-mentioned categories. Multi-word expressions or separable verbs were not included in the lists. The lists were composed on two native speaker's judgements with a good knowledge of the literature on perspectivisation.

The idea of the thematic text analysis was to confirm that the break in semantic tightness occurring with *want*-segments, as revealed by the LSA-analysis, could indeed be interpreted in terms of a perspective shift. We would therefore expect that the causal segments related by the connective *want* show diverging perspectivisation patterns, and that this will not be the case for the segments related by *omdat, doordat, aangezien.* This is reformulated in hypothesis 3.

> Hypothesis 3: If the causal segments are related by the connective *want*, the Q-segment contains perspective signals, the P-segment does not. The causal segments related by the connectives *omdat, doordat, aangezien* do not present such a shift.

| | Communication markers | | Attitude markers | |
|---|---|---|---|---|
| | Mean Q | Mean P | Mean Q | Mean P |
| aangezien (N = 139) | 0.173<br>SD: 0.38 | 0.115<br>SD: 0.32 | 0.360<br>SD: 0.48 | 0.273<br>SD: 0.45 |
| doordat (N = 699) | 0.129<br>SD: 0.33 | 0.104<br>SD: 0.31 | 0.305<br>SD: 0.46 | 0.326<br>SD: 0.47 |
| omdat (N = 6747) | 0.179<br>SD: 0.38 | 0.162<br>SD: 0.37 | 0.312<br>SD: 0.46 | 0.312<br>SD: 0.46 |
| want (N = 5589) | 0.175<br>SD: 0.38 | 0.181<br>SD: 0.38 | 0.442<br>SD: 0.50 | 0.394<br>SD: 0.49 |

Table 3: Mean number of perspective markers in P & Q

The results displayed in Table 3 show that the hypothesis is borne out for the subcategory of attitudinal markers: *want*-segments display a higher amount of attitudinal markers in Q than in P (F(1, 5588) = 26.84, p < 0.0001). For the other connectives this is not the case. For the communication markers, the hypothesis is not borne out. Actually, only *omdat* displays a higher amount of communication markers in Q (F(1, 6746) = 6.53, $p < 0.01$). While this latter result might seem counter to expectation, it actually goes in the direction of prior observations that *omdat*-relations frequently display the explicit introduction of speech acts (Degand, 2001; Pit 2003).

All together, these results offer new interesting insights into the discourse environment of (Dutch) causal connectives. On the one hand, we have shown with the LSA analysis that the proximity between Q and P is lower for *want*-relations than for the other connectives and that this is also the case for the semantic proximity between the sentences prior and subsequent to the causal relations. We therefore concluded that the connective *want* is a marker of thematic shift. On the other hand, the TTA analysis revealed that the Q-segments in *want*-relations display a higher amount of attitudinal markers. In our view, the presence of these markers leads to the conclusion that the connective *want* is indeed a marker of perspective shift, i.e. the break in semantic tightness should be interpreted as a perspective break, as has often been suggested in the literature. Furthermore, the additional results for *want* (absence of communication markers in Q) also suggest that markers expressing the speaker's attitude should be clearly distinguished from those that explicit the speaker's speech act (verbs of saying) or designate him/her explicitly as the source of the speech act (adverbs like *aldus*, *volgens*, … 'according to').

The polyphony/monophony distinction overlaps with the coordination/subordination distinction between *want* vs. the other connectives. The question arises which of those two factors is responsible for the results obtained. One route to follow is to compare our results with a language like English in which a same connective (*because*) has both monophonic and polyphonic uses, or with a language like French where a polyphonic connective like *puisque* is subordinating. The latter topic is object of ongoing research.

## 4 Discussion

In this paper we have presented a method for the linguistic investigation of a discourse phenomenon, viz. connectives, giving very satisfying results without necessitating heavy, work-intensive (hand-based) discourse annotation. The research presented is important to the corpus study of discourse phenomena for a number of reasons. The first is that it makes it possible to test linguistic hypotheses about the use of causal connectives on a large scale basis, whereas previous tests were based on only small corpora and small amount of data. The second is that the analysis is mostly fully automatic, especially with respect to the coding of the fragments. It is especially this latter feature that should appeal to the linguistic community, and makes our method more robust. The intercoder reliability is a constant concern of everyone working with corpora to test linguistic hypotheses (Carletta, 1996), and the more so when one is coding for semanto-pragmatic interpretations, as in the case of the analysis of connectives. A third reason is that our method combines two techniques of automatic text analysis, which allows us to formulate our hypotheses to be tested more fine-grained than possible with either one separately. Moreover, hypothesis formulation and testing goes further: We can use the methodology to formulate new hypotheses. An interesting possibility is to use LSA to find neighbours of terms in the dictionary, thus extending the dictionary. A further interesting venue is to test the linguistic hypotheses for different genres. This brings us to a further possibility, namely to reuse the semantic space for different types of linguistic research. A final possibility is to use the present semantic space for comparative research: How do the present results compare to a similar analysis of French connectives?

## Acknowledgements

L. Degand and Y. Bestgen are research fellows of the Belgian National Fund for Scientific Research. This research was supported by grant n° FRFC 2.4535.02 and by a grant (“Action de Recherche concertée”) of the government of the French-language community of Belgium.

## References

Berry, M.W. (1992). Large scale singular value computation, *International journal of Supercomputer Application*, 6: 13-49.

Berry, M., Do, T., O'Brien, G., Krishna, V. and Varadhan, S. (1993). *SVDPACKC: Version* 1.0 *User's Guide*, Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.

Burgess C., Livesay K., Lund K., " Explorations in Context Space : Words, Sentences, Discourse ", *Discourse Processes*, Vol. 25, 1998, p. 211-257.

Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. *Actes du Colloque International sur la Fouille de Texte CIFT'02* (pp. 81-94). Nancy : INRIA.

Bestgen, Y., Degand, L. & Spooren, W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an exploratory study. Lagerwerf L., Spooren W., Degand L. (Eds). *Determination of Information and Tenor in Texts: MAD 2003,* Stichting Neerlandistiek VU Amsterdam & Nodus Publikationen Münster, 179-188.

Biber, D. (1998). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Brouwers, L. (1997). *Het juiste woord, betekeniswoordenboek*. 6th ed. (ed. by F. Claes). Antwerpen etc.: Standaard.

Caenepeel, M. (1989). *Aspect, Temporal Ordering and Perspective in Narrative Fiction*. Doctoral Dissertation University of Edinburgh.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics 22* (2), 249-254.

Choi, F., Wiemer-Hastings P., & Moore J. (2001) Latent Semantic Analysis for Text Segmentation. In L. Lee & D. Harman (Eds.), Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing , 109-117.

Daelemans, W., Zavrel, J., Berck, P., & Gillis, S. (1996). MBT: A Memory-Based Part of Speech Tagger-Generator. In E. Ejerhed & I. Dagan (Eds.), *Proceedings of the Fourth Workshop on Very Large Corpora* (pp. 14-27). Copenhagen, Denmark.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R., Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, 1990, 391-407.

Degand, L. (2001). *Form and Function of Causation. A theoretical and empirical investigation of causal constructions in Dutch,* Peeters, Leuven, Paris, Sterling.

Degand, L. & Pander Maat, H. (2003) A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale, A. Verhagen & J. van de Weijer (eds.) *Usage based approaches to Dutch* (pp. 175-199). Utrecht: LOT.

Foltz, P.W., Kintsch, W., & Landauer T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review, 7,* 257-266.

Kintsch W., (2001).Predication, *Cognitive Science* 25, 173-202.

Landauer, T.K., Foltz, P.W., and Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, *25* (2, 3), 259-284.

Lebart, L., Salem, A., and Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Publisher.

Lemaire, B., Bianco, M., Sylvestre, E., & Noveck, I. (2001). Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente. In H. Paugam Moisy, V. Nyckees, J. Caron-Pargue (Eds.), *La Cognition entre Individu et Société : Actes du Colloque de l'ARCo* (pp. 309-320). Paris: Hermès.

Oversteegen, L. (1997). On the pragmatic nature of causal and contrastive connectives. *Discourse Processes,* 24, 51-86.

Pennebaker, J.W., , Mehl, M.R., & Niederhoffer, K.G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577.

Piérard, S., Degand, L., & Bestgen Y. (2004). Vers une recherche automatique des marqueurs de la segmentation du discours. *Actes des 7$^{es}$ Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve.

Pit, M. (2003). *How to Express Yourself with a Causal Connective. Subjectivity and Causal Connectives in Dutch, German and French*. Amsterdam : Rodopi.

Popping, R. (2000). Computer-assisted text analysis. London: SAGE.

Spooren, W.P.M.S (1989). *Some Aspects of the Form and Interpretation of Global Contrastive Coherence Relations.* Unpublished Dissertation, K.U. Nijmegen.

Stone, P.J. (1997). Thematic text analysis: New agendas for analyzing text content. In C.W. Roberts (Eds.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (pp.35-54). Mahwah, NJ: Erlbaum.

van den Bosch, A., & Daelemans, W. (1999). Memory-based morphological analysis. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, ACL'99 (pp. 285-292). New Brunswick, NJ: ACL