Studying the Semantic Context of two Dutch Causal Connectives

Iris Hendrickx and **Wilbert Spooren**
Centre for Language Studies, Radboud University Nijmegen
P.O. Box 9103, NL-6500 HD Nijmegen The Netherlands
i.hendrickx, w.spooren@let.ru.nl

Abstract

We aim to study the difference of usage between two causal connectives in their semantic context. We present an ongoing study of two Dutch backward causal connectives *omdat* and *want*. Previous linguistic research has shown that causal constructions with *want* are more subjective and often express an opinion. Our hypothesis is that the left and right context surrounding the connectives are more semantically similar in sentences with *omdat* than sentences with *want*. To test this hypothesis we apply two techniques, Latent Semantic Analysis and n-gram overlap. We show that both methods indeed indicate a substantial difference between the two connectives but opposite to what we had expected.

1 Introduction

Much corpus linguistic research has dealt with the issue of subjectivity, i.e. the degree to which the presence of the writer or speaker of a text is felt ((Sanders and Spooren, 2013), and the references cited there). Subjectivity can be located at different levels in a text. At the word level, some words (e.g., evaluative adjectives and expletives) imply a writer/speaker evaluation, whereas others do not. At the sentence level, the description of facts is felt to be more objective, whereas opinions are more subjective. And at the supra-sentential level, subjectivity can get expressed in the type of relation that links the clauses or sentences. For example, argumentative relations are more subjective than statements. Interestingly, many languages make a distinction between more objective or more subjective causal connectives. In Dutch, for example, *omdat* is typically used to express more or less objective backward causal relations, whereas *want* is typically used for more subjective relations. However, these connectives are near synonyms and can be used in the same context as shown in example 1 and 2. There is subtle difference in meaning because example 1 focuses on the reason relation between the two segments whereas 2 focuses on the argument relation. As the first segment is an opinion, *want* is slightly more natural than *omdat*.

(1) Dat is vooral jammer **omdat** de hoofdrolspeler uitstekend zingt.

(2) Dat is vooral jammer **want** de hoofdrolspeler zingt uitstekend.
"That is particularly unfortunate because the protagonist sings excellent."

Note the difference in word order: *want* leads to a coordinative conjunction while *omdat* gives a subordinate conjunction.

We need more insight into this subtle difference between connectives for example to allow natural language generation systems to mimic the choices that native speakers of Dutch make intuitively. Another application would be sentiment analysis where the difference in subjectivity of various connectives can be used to identify subjective or opinionated sentences.

Presently the corpus linguistic analyses of subjective versus objective causal relations have very much been a small-scale enterprise, in that corpus examples were annotated manually. This is problematic for at least two reasons: manual annotation relies on hand coding, with the accompanying problems of poor inter-annotator reliability, and the restricted size of the hand annotated corpora limits the power of statistical generalization. Bestgen et al. (2006) suggested to complement these manual analyses with automatic analyses.

Bestgen and colleagues studied backward causal connections in Dutch. They made use of

Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language, pages 28–32,
Gothenburg, Sweden, April 26, 2014.

two types of automatic analyses: Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and what they call Thematic Text Analysis (Popping, 2000) to show that the semantic connection between first and second segment is weaker in a *want* connection than in a *omdat* connection, and that the first segment of *want* connections contains more subjective words than the first segment of *omdat* connections. The materials that were used by Bestgen et al. (2006) were texts from a large corpus of newspaper language of 16.5 million tokens.

The purpose of our current ongoing research project is to extend the automatic analyses in two ways: on the one hand we want to reproduce the LSA analysis of Bestgen et al. using a larger corpus of about 30 million tokens; on the other hand, we want to use n-gram analyses to investigate the semantic connection between the segments in a *want* versus *omdat* connection.

The use of n-grams to measure semantic overlap is a well known method, which has been applied in the standard evaluation metrics for tasks like machine translation and automatic summarization. In these tasks automatic systems aim to produce a text as similar as possible to a manually constructed gold standard text. To evaluate the quality of these automatically produced text, measures such as BLUE (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003) measure n-gram overlap between the system text and the gold standard text. Furthermore, in other types of research like in the field of literary studies n-grams have been applied, for example to discriminate between genres (Louwerse et al., 2008) or for author discrimination (Hirst and Feiguina, 2007).

Backward causal connectives denotes a cause relation. The connective is positioned in a sentence between the consequence (denoted as Q) and the cause (denoted as P). For the sentence in example 1 Q is the text segment before the connective, and P contains all words after the connective as follows:

Q Dat is vooral jammer

P de hoofdrolspeler uitstekend zingt.

Our hypothesis is that Q and P are more semantically similar in sentences with *omdat* than sentences with *want*. This implies that we expect the average cosine between P and Q to be smaller in *omdat* connections than in *want* connections. We also hypothesize that the number of n-grams shared between P and Q will be higher in *omdat* sentences than in *want* sentences.

This paper presents work in progress. We first describe the SoNaR corpus that was used in this study in section 2. In section 3 we present the experimental setup and results of the experiments with LSA. In section 4 we detail our approach to computing n-grams and we discuss our findings and the next steps to take in 5.

2 Data Collection

Unfortunately neither the corpus nor the data sample used by Bestgen et al. (2006) was available to us. For this reason we chose a similar Duch corpus to work with. The SoNaR corpus (Oostdijk et al., 2013) is a reference corpus of 500 million written words of contemporary Dutch sampled from a wide variety of sources and genres. The corpus has been automatically tokenized, part-of-speech tagged and lemmatized. We took a sample of 100K news articles from the SoNaR corpus as our experimental data set. As we are interested in semantic overlap, we took the lemmatized versions of the articles.

From this data set, we collected all sentences containing the connectives *omdat* and *want*. As we aim to study the semantic relation between Q and P, we only selected sentences that have a meaningful Q and P in the same sentence. We excluded sentences with sentence initial connectives as they only contain a P segment. Sentences with short Q segments (containing one or two words), were manually inspected. A sentence that starts with *dat komt omdat* " this is because" does not contain a meaningful consequence because it refers back to information in a previous sentence. On the other hand, a short Q segment like *tevergeefs, want* "in vain, because" does express a meaningful consequence. In case of sentences with multiple connections, we took the first Q and P and cut off the remainder parts using some handwritten rules. Overall we excluded 20% of *want* sentences and 25% of *omdat* sentences. In total we selected 18,260 for *omdat* and 14,449 sentences for *want*. Some statistics about the sentences is shown in Table 1.

3 LSA

Latent Semantic Analysis (LSA) is a mathematical method for representing word meaning similarity in a semantic space based on a term-by-documents

	Sentences	length	Q len	P len
omdat	18,260	24.3	11.2	12.1
want	14,449	23.5	9.6	12.9

Table 1: Number of sentences and average length in tokens of the full sentence, Q, and P in the data set of *want* and *omdat*.

matrix. It applies singular value decomposition to this matrix to condense it to a smaller semantic representation of around 100 - 500 dimensions (Landauer et al., 1998).

We applied LSA to measure the semantic overlap between Q and P of the *omdat* and *want* sentences. We constructed a term-by-document matrix based on the SoNaR news sample and converted this to an LSA space with 300 dimensions. Each Q and P was projected as a term vector in the LSA space and we computed the cosine similarity between each Q and P.

To build the document-by-term matrix for LSA, words were lemmatized, and punctuation, digits and stopwords (based on a stopword list of 221 words) were filtered out.

In our first analysis we used the top most frequent words that occurred at least 15 times, leading to a text matrix of approximately 20,000 documents and 19,000 word terms. We calculated the cosine between Q and P for each of the *omdat* and *want* sequences. A Welch Two Sample t-test showed that contrary to expectation the cosine between Q and P was lower for *omdat* (0.039) than for *want* (0.045; $t(29518)$=-4.78, p <.001).

In a second analysis we chose a sample of a different scale and we used a text matrix of 100,000 documents and the top 10,000 most frequent word terms. A t-test showed that in this case the cosine for *omdat* sequences was slightly but significantly higher than for *want* sequences (*omdat*: 0.048; *want*: 0.043; $t(30175)$=3.68, p <.001).

In the final section we will go into possible explanations for these unexpected and incompatible results.

4 N-gram overlap

In our study of n-grams, we looked both at pure bigram statistics and at n-grams in a broader scope, i.e. n-grams and skip-grams with a maximal length of 10 tokens. All n-grams have a minimum length of 2, and a minimum frequency of 2 in the datasample. We use lemmatized words to reduce the influence of morphological information. For the n-gram analysis we used the Colibri software package developed by Maarten van Gompel[1] (van Gompel, 2014). In the left part of Table 2 we show the bigram statistics and on the right side the n-gram statistics of n-grams that occur at least twice in Q, P, and those occurring in both Q and P. We present the following counts:

- Pattern - The number of distinct n-gram patterns (n-gram type count)
- Coverage - The number of unigram word tokens covered as a fraction of the total number of unigram tokens.
- Occurrences - Cumulative occurrence count of all the patterns (n-gram token count).

We can observe that about 75% of the tokens in Q and P is covered in this bigram analysis, while the n-grams cover around 93% of the words. Zooming in on the bigrams and n-grams that are shared in Q and P, we can see that these cover about 50% and 75% of the tokens respectively. This shows that we can safely discard n-grams that occur only once in our counts and still cover most tokens in the data sample.

Based on the bigram occurrences in our data set, we computed whether the bigram overlap between Q and P in *omdat* sentences is larger than in *want* sentences. We used a loglikelihood test to compare the relative frequencies as our samples do not have the same size. We found that 72362 bigram occurrences (or 67.8%) overlap in *omdat* sentences and 58213 bigrams (or 79.4%) for *want* sentences (LL2(1)=808.40, p <.01). This means that, contrary to our hypothesis, we found more overlap for *want* sentences.

We performed the same computation on the larger set of n-grams. We saw that 81573 of n-gram occurrences (44.9%) overlap in *omdat* sentences and 65272 (51.1%) overlap in for *want* sentences (LL2(1)=595.37, p <.01). This then is again a confirmation that we find more overlap between Q and P in *want* sentences.

5 Conclusions

In this paper we report two types of automatic analyses of the differences between *want* and *om-*

[1] available at: http://proycon.github.io/colibri-core/

Category	Bigrams			n-grams		
	Patterns	**Coverage**	**Occurrences**	**Patterns**	**Coverage**	**Occurrences**
omdat Q	18931	0.7312	106766	39780	0.9320	181506
omdat P	20649	0.7549	118414	45074	0.9380	208809
omdat Q&P	7261	0.5042	72362	9213	0.8927	81573
want Q	12938	0.7474	73276	27654	0.9350	127723
want P	17564	0.7216	94685	37027	0.9271	159125
want Q&P	5774	0.4847	58213	7365	0.7943	65272

Table 2: Counts of the bigrams and n-grams up to length 10 with minimal frequency 2 in Q, P, and those n-grams that occur in both Q and P. Patterns refers to n-gram types, Occurrences to n-gram tokens and Coverage refers to word token coverage.

dat, which have been claimed to differ in subjectivity, i.e. the degree to which the writer is felt present in the text. One part of our study is a reproduction of (Bestgen et al., 2006) and assessed the semantic relationship between Q and P in terms of a LSA cosine for *want* and *omdat*. Contrary to the findings of Bestgen et al., our first LSA analysis showed that the relationship between Q and P is less strong for *omdat* than for *want*. A second analysis found a small difference in the expected direction. In the second part of our study we used n-gram overlap as a different type of similarity measure. Again, our hypothesis was not borne out in that *omdat* showed a significantly smaller degree of overlap than *want*.

At this moment we cannot explain why the two LSA experiments presented in section 3 show significant results in different directions. In the two experiments the same connective sentences were used, but the semantic space in which they were projected was different. For our LSA analysis we made use of the software package LSA in R. To rule out the possibility that our results were due to some implementation peculiarity, we ran a small test sample with another LSA implementation Gensim (Řehůřek and Sojka, 2010). Both implementations gave us similar cosine values for the same sample.

A noticeable difference with the Bestgen et al. study is the size of the cosines: Bestgen et al. report mean cosines of 0.120 and 0.137 for *want* and *omdat*, respectively, whereas in our study we found mean cosines of 0.045 and 0.039, respectively. This suggests that our data sample and experimental setup differ substantially from the work of Bestgen et al. and we did not succeed in reproducing their experiment. In our analysis the semantic relationship between Q and P is much weaker.

In order to be able to interpret these results, we added a baseline experiment. Here we ran an LSA experiment with segments composed of random words of the exact same size for the *omdat* and *want* sentences. For *omdat* this gave us a mean cosine similarity of 0.007 and for *want* 0.006. This implies that the cosines we found are significantly higher than comparing random strings of words.

Note that the analysis was carried out on a sufficiently large corpus and sufficient numbers of occurrences of *want* and *omdat*. Moreover, the result that semantic relationship is stronger in *want* than in *omdat* is corroborated by our n-gram analysis.

One possible explanation of the results of the n-gram analysis is the syntactic difference between *want* and *omdat* sentences. In *want* sentences the word order of Q and P is the same while for *omdat* the verb-predicate order is swapped. The n-grams will pick up this difference. As a next step we plan to run the n-gram analysis with alphabetically ordered n-grams to exclude the effect of this syntactic difference[2].

Another line of future research is to make genre comparisons. The availability of the SoNaR corpus makes it possible to investigate the subjectivity hypothesis for different text genres.

Finally we intend to follow up our analysis with a machine learning experiment to investigate whether a learner could distinguish a *want* sentence from a *omdat* sentence by looking at a local context window of words to automatically predict *want* or *omdat*.

[2]We wish to thank one of our anonymous reviewers for bringing this suggestion to our attention.

References

Yves Bestgen, Liesbeth Degand, and Wilbert Spooren. 2006. Toward automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes*, 41(2):175–193.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Graeme Hirst and Olga Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

C.-Y. Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 71 – 78, Edmonton, Canada.

Max Louwerse, Nick Benesh, and Bin Zhang, 2008. *Directions in Empirical Literary Studies: In honor of Willie van Peer*, chapter Computationally discriminating literary from non-literary texts, pages 175–191. John Benjamins Publishing.

Nelleke Oostdijk, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written dutch. In *Essential Speech and Language Technology for Dutch*, pages 219–247. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. pages 311–318.

R. Popping. 2000. *Computer-assisted text analysis*. Sage, London.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

T.J.M. Sanders and W.P.M.S. Spooren. 2013. Exceptions to rules: a qualitative analysis of backward causal connectives in Dutch naturalistic discourse. *Text & Talk*, 33(3):399–420.

Maarten van Gompel, 2014. *Colibri Documentation, Colibri Core 0.1*. Centre for Language Studies, Radboud University Nijmegen, The Netherlands. http://proycon.github.io/colibri-core/doc/.