Cueing the Virtual Storyteller: Analysis of cue phrase usage in fairy tales

Manon Penning and **Mariët Theune**
Human Media Interaction
University of Twente
Enschede, The Netherlands
m.p.j.penning@student.utwente.nl, m.theune@utwente.nl

Abstract

An existing taxonomy of Dutch cue phrases, designed for use in story generation, was validated by analysing cue phrase usage in a corpus of classical fairy tales. The analysis led to some adaptations of the original taxonomy.

1 Introduction

A taxonomy of Dutch cue phrases, used to signal rhetorical relations between text segments, has been developed for the generation of narratives in the Virtual Storyteller project (Theune et al., 2006; Theune et al., 2007).[1] The taxonomy includes only the most frequent cue phrases found in the Spoken Dutch Corpus.[2] Because the Spoken Dutch Corpus consists largely of spontaneous speech, the taxonomy might not be fully representative of cue phrase usage in the target domain of the Virtual Storyteller, which is fairy tales. In this paper we describe a corpus analysis we carried out to investigate this issue, and we discuss the modifications we made to the taxonomy based on the results. We also present a preliminary comparison of cue phrase usage in direct and indirect discourse in fairy tales.

2 The corpus

The Dutch "Stichting Beleven" has a large on line collection of Dutch translations of classical fairy tales and fables (Stichting Beleven, 2006). They tried to collect translations that are as true to the stories in the original tales as possible, while avoiding archaic language. Therefore, we considered this website to be a useful source for the purpose of this research. From the website we selected 8 fairy tales by Aesop, 25 by Andersen and 25 by the brothers Grimm. In the case of Aesop, this included all available stories by this author. In the cases of Andersen and Grimm, selections were made based on the popularity of the stories on the website. This resulted in a corpus of 97.000 words.

3 Procedure

The goal of our analysis was to find out whether the 36 cue phrases in the taxonomy of Theune et al. (2006) were in fact among those most frequently used in fairy tales. We also wanted to find any cue phrases that did appear in fairy tales but were not in the taxonomy. To identify potential cue phrases we first collected a list of all unique words occurring in the fairy tale corpus, and then determined for each word whether it could be used as a cue phrase by formulating one or more sentences in which the word occurred as a cue phrase. This resulted in a list of 85 potential cue phrases of which we wanted to determine the frequency in the corpus. A complicating factor here was that words that are used as cue phrases can sometimes also have a different function. Litman (1996) has labeled these two types of occurrences with 'discourse sense' (when actually used as a cue phrase) and 'sentential sense' (when used as some sort of filler, noun, verb or other non-cue phrase type of word). For example the Dutch word 'maar' (but) can be used to indicate contrast as in "Zij hadden mooie blanke gezichtjes, *maar* ze waren lelijk en zwart van hart." (They had beautiful white faces, but their hearts were ugly and black), but also as some sort of filler "Wacht *maar*, ik krijg je nog wel!" (Just wait, I'm gonna get you yet!).

Indicators whether a potential cue phrase is used in its sentential or discourse sense include part of speech, the presence of collocations and ortho-

[1] http://wwwhome.cs.utwente.nl/~theune/VS/index.html
[2] http://lands.let.kun.nl/cgn/ehome.htm

graphic markers, and the position of the cue phrase in the utterance (Hirschberg and Litman, 1994; Litman, 1996; Oates, 2000; Louwerse and Mitchell, 2003; Zufferey and Popescu-Belis, 2004). For the potential cue phrases in our corpus, we determined manually for each occurrence whether it was a case of discourse or sentential use.[3] In the case of discourse use, it was determined which relation from the taxonomy was signalled: causal, additive, contrastive or temporal (or more specific subtypes of those relations), or possibly another relation not included in the taxonomy. This was done largely based on intuition, but when in doubt we used a variant of the substitutability test of Knott and Dale (1994), allowing all substitutions that did not influence the meaning of the sentence. This was done independently by two annotators. We did not measure annotator agreement, but only compared the resulting classifications, resolving any differences through discussion. A few uncertain cases remained, so this procedure did not result in exact counts, but it was sufficient to get a general idea of how often the various cue phrases were used.[4]

4 Results and adaptation of the taxonomy

Table 1 gives an overview of the frequencies of the cue phrases in our corpus. The cue phrases from the original taxonomy are given in italics. As can be seen in Table 1, there are a lot of cue phrases that only occur rarely and a few cue phrases that occur quite often. Temporal relations seem to be signalled much less often than additive, cause and contrast relations. Some of the words that had been identified as potential cue phrases did not occur as actual cue phrases in the corpus at all (#cue = 0). This category also included a few of the cue phrases from the original taxonomy. The table also shows that some of the newly identified cue phrases seem to be good alternatives for less used ones that already were in the taxonomy. For example, the word 'toen' (then) is a temporal marker that was not in the original taxonomy but occurs very frequently in fairy tales.

The analysis did not give rise to adaptations of the structure of the taxonomy, because most of the cue phrases found in the corpus could be easily fitted into the existing relation (sub)categories. One exception was the cue phrase 'anders' (otherwise) which signalled an 'Otherwise' relation not in the taxonomy. However, we decided not to add it to the taxonomy because our generation system currently cannot produce this type of relation (Theune et al., 2007). We also added a new category for negative additives ('evenmin' and 'noch', meaning something like 'neither'), but we did not add these to the taxonomy because their counts were very low and 'noch' in particular is a bit archaic.

All in all, we kept the structure of the original taxonomy as it was, but we did make some changes in the cue phrases included in the taxonomy. For a start, the cue phrases that did not occur in the corpus at all were removed ('ooit', 'uiteindelijk', 'vervolgens' and 'waardoor'). Secondly, some of the cue phrases that did not occur very often and did not seem to differ in meaning from other, more frequent alternatives were removed: 'en…ook', 'zowel…als' (additive), 'en' (causal) and 'doordat' (involuntary cause-last).[5] Also, we replaced the less frequent cue phrase 'plotseling' (suddenly) by the more frequent synonym 'opeens'. Based on the high counts of 'eerst' (first) and 'toen' (then), it was decided to add those to the taxonomy. The cue phrases that we kept in, or added to, the new version of the taxonomy are shown in bold face in Table 1.

5 Direct vs. indirect discourse

It has been noted that cue phrase usage differs between monologues and dialogues (Louwerse and Mitchell, 2003). Since in addition to just descriptive, indirect discourse, fairy tales tend to have pieces of direct speech in them (e.g., "What big ears you have, grandma"), we carried out an additional small-scale investigation to find out if there were any differences between those two text types in our corpus. For this study we selected 20 fairy tales that contained at least 5 lines of direct discourse and split them into collections of direct and indirect discourse (8.493 and 24.967 words respectively).

The cue phrase frequencies in these collections are summarised in Table 2. Because we had about three times as much data for indirect discourse, for a fair comparison we used relative counts here (number of occurrences every 10.000 words) instead of

[3] Implementing precise rules to automatically distinguish between these cases would have been more time consuming. Also, the use of different character encodings in the corpus hindered automatic processing.

[4] Only for the potential cue word 'en' another procedure was used: since it occurred over 4000 times in our corpus, checking each instance by hand was infeasible. Therefore, in this case we extrapolated from a number of random samples.

[5] We regard the (equally frequent) cause-first version of 'doordat' as the preferred alternative, because mentioning the cause first makes the generated stories easier to read.

Relation	**Primitive**		#cue = 0	0 < #cue ≤ 10	10 < #cue ≤ 50	50 < #cue ≤ 100	#cue > 100
Cause	voluntary	cause-first	hierdoor, vandaar	zodoende	***daarom, dus***, ***omdat***, dan ook		
		cause-last		tenslotte	immers		***want***
		purpose		ervoor	opdat, zodat		***om***
	involuntary	cause-first	*waardoor*	***daardoor***, ***doordat***, dus	***zodat***		
		cause-last		*doordat*			
					en, ***omdat***		
Additive	moreover			***bovendien***, daarbij, ook nog	zelfs		
				alsmede, daarbij, verder, evenals, *zowel . . . als*	*en. . . ook*		***en***, ***ook***
	negative			evenmin, noch			
Contrast	unrealized cause		evengoed	daarentegen, evenwel, ***hoewel***, niettegenstaande dat, ofschoon, ondanks, ook al, weliswaar			***toch***
				alleen	***echter***		***maar***
Temporal	after	gap	*ooit*	***later***			
		sequence	*vervolgens*	***nadat***, sinds, sindsdien, straks, vanaf, waarop	***daarna***, na		**toen**
	before	gap	*ooit*	eerder, laatst, ***vroeger***			
		sequence		daarvoor, eer, ***voordat***, tevoren, totdat	**eerst**		
	finally		*uiteindelijk*	op het laatst	***eindelijk***, ***tenslotte***		
	suddenly			ineens, *plotseling*	**opeens**		
	during		gedurende, tussendoor	daarbij, in de tussentijd, ondertussen, intussen, onderwijl, zolang	***terwijl***		
	once				***eens***		
	when	as soon as			***zodra***		
						wanneer	**toen**, ***als***
Other					anders		

Table 1: Counts of cue phrases (#cue) organised by the relations they signal (based on (Theune et al., 2006)). Cue phrases from the original taxonomy are shown in italics; cue phrases included in the adapted taxonomy are shown in bold face.

Indirect discourse	Direct discourse			
	#cue = 0	0 < #cue ≤ 5	5 < #cue ≤ 20	#cue > 20
#cue = 0	alleen, daarvoor, eer, in de tussentijd, ineens, laatst, niettegenstaande dat, noch, ondertussen, straks, tevoren, totdat, uiteindelijk, vanaf, vroeger, waarop, zodoende	daarentegen, eerder, ervoor, ofschoon, onderwijl, zowel. . . als		
0 < #cue ≤ 5	alsmede, bovendien, daarbij, daardoor, evenals, evenmin, evenwel, in de tussentijd, intussen, later, ondanks, opeens, op het laatst, plotseling, sinds, sindsdien, verder, weliswaar, zodra, zolang	anders, doordat, dus, echter, eerst, en. . . ook, hoewel, immers, na, nadat, opdat, tenslotte, voordat	wanneer	
5 < #cue ≤ 20	ondanks	daarna, daarom, eindelijk, eens, omdat, terwijl, zelfs	als, zodat	toch
#cue > 20		toen	om, want	en,maar, ook

Table 2: Cross-table for counts of cue phrases (#cue) in direct and indirect discourse (per 10.000 words).

absolute numbers. Since the total number of cue phrases was much smaller than in the full corpus, the ranges used in Table 2 were adapted accordingly. The cue phrases with #cue < 0 in Table 1 were left out since they would only meaninglessly clutter up the table. Still, quite a number of cue phrases that did appear in the total collection of fairy tales, did not occur in the selection of 20 fairy tales used here.

Table 2 shows that the most frequent cue phrases from the overall collection, also occur most often in both direct and indirect discourse. An exception is 'omdat' (because), which occurs more often in indirect than direct discourse. This is consistent with research by Degand and Pander Maat (2003) showing that the alternative 'want' is preferred when the speaker is somehow personally involved with the action being described (which is more typically the case in direct speech). Furthermore, in indirect discourse more cue phrases are used than in direct discourse. Responsible for this difference are mostly the less common causal cue phrases and temporal cue phrases. This can be explained intuitively by the difference in nature between direct and indirect discourse. When characters in a story engage in conversation, they are likely to discuss simple, current events without using elaborate language. But in narrating a story, a number of events is summed up mentioning actions, consequences, causes and temporal span. All in all, our findings are in line with earlier research, but the small number of data does not allow us to draw substantial conclusions.

6 Concluding remarks

We carried out an analysis of cue phrase usage in fairy tales in order to validate a cue phrase taxonomy for story generation in Dutch. This led to some modifications of the taxonomy such as leaving out the least frequent phrases and replacing others by more frequent alternatives. Limiting the taxonomy in this way to a small number of the most common cue phrases could make the generated stories easier to read (Williams and Reiter, 2005), but on the other hand it could also make them more boring (Knott and Dale, 1994). However, the adapted taxonomy seems to contain a sufficient number of alternatives for each cue phrase to limit the latter risk, as we hope to show during future evaluations.

Although differences in cue phrase usage between different text types are to be expected, the limited extent to which modifications to the taxonomy were necessary shows that the difference between cue phrase usage in the Spoken Dutch Corpus and cue phrase usage in classical fairy tales is quite small. This indicates that the taxonomy is usable for a broader scale of texts than just fairy tale-like stories. However, when comparing direct and indirect discourse in fairy tales, some differences surface that might indicate a need for different taxonomies for both kinds of discourse. A larger scale study is needed to further investigate this.

References

L. Degand and H. Pander Maat. 2003. A contrastive study of Dutch and French causal connectives on the speaker involvement scale. In A. Verhagen and J. van de Weijer, editors, *Usage based approaches to Dutch*, pages 175–199. LOT, Utrecht.

J Hirschberg and D. Litman. 1994. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

A. Knott and R. Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18(1):35–62.

D. J. Litman. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

M. M. Louwerse and H. H. Mitchell. 2003. Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse Processes*, 35(3):199–239.

S.L. Oates. 2000. Multiple discourse marker occurrence: Creating hierarchies for coherence relations. In *Proceedings of the ANLP-NAACL 2000 Workshop on Student Research*.

Stichting Beleven. 2006. Wereld volksverhalen almanak. Retrieved on 18-09-2006, from: http://www.beleven.org/verhalen/.

M. Theune, F. Hielkema, and P. Hendriks. 2006. Performing aggregation and ellipsis using discourse structures. *Research on Language and Computation*, 4(4):353–375.

M. Theune, N. Slabbers, and F. Hielkema. 2007. The Narrator: NLG for digital storytelling. These proceedings.

S. Williams and E. Reiter. 2005. Generating readable texts for readers with low basic skills. In *Proceedings of ENLG-05*, pages 140–147.

S. Zufferey and A. Popescu-Belis. 2004. Towards automatic identification of discourse markers in dialogs: the case of like. In *Proceedings of SIGdial 2004*, pages 63–71.