Casting a Spell: Identification and Ranking of Actors in Folktales

Folgert Karsdorp, Peter van Kranenburg, Theo Meder,
Antal van den Bosch
Email: {folgert.karsdorp,peter.van.kranenburg,
theo.meder}@meertens.knaw.nl
a.vandenbosch@let.ru.nl

Abstract

We present a system to extract ranked lists of actors from fairytales ordered by importance. This task requires more than a straightforward application of generic methods such as Named Entity Recognition. We show that by focusing on two specific linguistic constructions that reflect the intentionality of a subject, direct and indirect speech, we obtain a high-precision method to extract the cast of a story. The system we propose contains a new method based on the dispersion of terms to rank the different cast members on a scale of importance to the story.

1 Introducing the Problem

Today do I bake, tomorrow I brew,
The day after that the queen's child comes in;
And oh! I am glad that nobody knew
That the name I am called is Rumpelstiltskin!

So says the song of the little man in Brother Grimm's *Rumpelstiltskin* with which he accidentally reveals his name to the queen. Advances in the task of Named Entity Recognition (NER) make it possible for computer systems to recognize his name as well. NER seeks to locate and extract atomic textual units into a predefined set of categories (e.g. PERSON, LOCATION and ORGANIZATION).

Now that we can recognize *Rumpelstiltskin* as being the name of a person, how do we know that he is a character or actor in the story? Are all named entities within the category of PERSON in a story automatically actors? How about characters that have no proper name in a story, but are referred to by nominals or noun phrases? How do we know that they are part of the cast of a story?

Most NER systems are developed for the domain of non-fiction, including newspapers, manuals and so forth. Applications of NER in literary texts or folktales, however, have been discussed only marginally (see [13] for an exception).

Unfortunately, the knowledge gained from traditional NER systems is not easily transferred to the domain of fictional texts, because of several incongruences between both domains. Named entities in fairytales, for example, are represented in ways quite different from those in non-fiction. We find traditional proper nouns such as *Hansel* and *Gretel* but also noun phrases such as *the king* and *the big bad wolf* which, although expressed differently, have the same function as their proper noun counterparts. To make things even more complicated, inanimate entities in folktales may belong to the cast as well. A nice example of this can be found in the story of *The Fleeing Pancake*, in which a pancake acts as the protagonist who tries to escape from its predators.[1]

In this paper we present a method for extracting the cast from fictional texts. The cast consists of a number of actors which we, following Bal [1], define as "agents that perform actions" and have intentions. An action is seen as to cause or experience an event. Actors are thus not necessarily human. As claimed by structuralists such as Greimas, actors are distinguished from other entities by having an intention. They strive towards a goal or aspire a particular aim [1]. The identification of actors in a story therefore would require a method for detecting intentionality. We show that direct and indirect speech are effective indicators of intentionality. Extracting these constructions from a text gives us a high precision method for retrieving the actors that form the cast of a story.

Not all actors in a story are equally important. The scale of importance ranges from what is customary referred to as 'the hero' of a story to mere background actors whose actions have relatively little influence on the course of the story. The system we propose ranks the different cast members on a scale of importance to the story on the basis of their dispersion in the text. We show that the ranking of the system corresponds to a large extent with that of expert knowledge.

The outline of the paper is as follows. We begin with some background information about actors from the perspective of narratology and combine this with insights from linguistics in Section 2. We then describe the corpus used and the way the annotations of the corpus have been established. Section 4 presents the computational methods used in this study. In Section 5 we report on the empirical results for the proposed method. The last section offers our conclusions and directions for further research.

2 Backgrounds

2.1 Narratological Background

We ground our definition of an actor in the theory of narratology as developed by Bal [1] who borrows from the French structuralist Greimas [6]. The underlying presupposition in the proposal by Greimas is that human activity (thinking and action) targets some aim. Actors are no different and show some intentionality of

[1] See for example `http://www.pitt.edu/~dash/type2025.html#gander`

achieving this aim or goal. This could be the achievement of something pleasant, some object of desire (be it physical or non-physical) or something disagreeable. Natural language reveals this intentionality in particular verbs (e.g. *to fear*, *to wish*).

Bal distinguishes six classes of actors. There are actors who pursue some aim (called SUBJECT) and the aim itself (called OBJECT). A SUBJECT x strives towards some aim y where y is an OBJECT. In a prototypical fairytale love story these roles might be filled with 'the prince who wants to marry the princess'. OBJECTS are not necessarily human or even animate but could also be a mental state or an aspiration such as 'becoming king'. Other types of actors in Bal's theory are POWER and RECEIVER. POWER is often an abstraction that supports the SUBJECT in reaching its goal or prevents it from getting there. The RECEIVER is the one to whom some desired object is 'given'. These two actor types are closely related to another pair of actors, namely the HELPER and the OPPONENT. As their name reveals, HELPERS help the SUBJECT in achieving its goal (although providing only incidental help) when some OPPONENT intends to distract it from doing so.

Actors are thus not necessarily human or animate beings, but can be of a more abstract nature as well, such as skills and abilities. In this paper we are only interested in the classes of actors that show intentions in the context of a particular story. Whether they are animate or not is only partly based on reality and is determined mainly in relation to the story itself.

2.2 Linguistic Background

A baseline approach to the problem would be to tag all animate nouns as actors. However, this compromises both on precision and recall, because not all animate entities are actors in a story and not all inanimate entities are not actors.[2] Following [1], we claim it is better to look for entities that exhibit intentionality or consciousness. Are there any linguistic clues for intentionality?

Sentences in which the intentionality of some entity is most clearly revealed are sentences containing direct speech. In traditional grammar the quoted clauses in the examples below are commonly characterized as (direct) objects of the verbs:

(1) *"Lay your head on the chopping block," said the old witch.*

(2) *"I heard the world cracking and I flee away," answered the goat.*

In sentences like these we read the text as if it was uttered by the actor. That we should do so is both linguistically (by means of syntax) and orthographically (by means of quotes) marked. However, speech can also be indirectly narrated, just as thoughts, as the following examples show:

(3) *The man asked why the old woman smashed all her eggs to pieces.*

[2] Moreover, the identification of animate nouns would require a semantic lexicon, which is a rather expensive resource not available for all languages.

(4) *The robbers thought the farmer had nothing to hide.*

In these examples of indirect speech it is less clear that the presented text has to be attributed to the actor; it may by influenced by the narrator's stance towards the situation. Traditionally, these sentences are characterized as complement constructions in which the subordinate clauses fill the role of (direct) objects of the verbs of the main clause. Most verbs that take such 'direct object clauses' belong to one of the following semantic types [7]:

i Verbs that express a command, question, promise, etc (i.e. verbs of communication such as *to tell, to promise, to beg*);

ii Verbs that denote knowing, believing, supposing, etc (*to imagine, to understand, to realize*);

iii Verbs that indicate the evaluation of something (*to appreciate, to regret*);

iv Verbs that express a wish or desire (*to wish, to hope*);

v Verbs that denote a particular way of perception (*to hear, to discover, to see*).

Verhagen points at an interesting commonality between these five verb types:

> "[...] such predicates all evoke a mental state or process of a subject of consciousness (sometimes a process *comprising* a mental state, as in the case of communication), and the content of the complement is associated with this subject's consciousness in a particular manner." [15, 100]

The resemblance with the idea of Greimas that actors are intentional entities aspiring a particular goal or aim, should be clear. If we can automatically retrieve the subject of the main clause of a complement construction, we have a good chance of finding an actor in a story. Both instances of direct speech and indirect speech provide us with cues about the intentionality of entities without the need to know the verb's semantics.

2.3 Computational Background

Few studies have focused on the recognition of actors in fictional texts. Declerck *et al.* [2] present an ontology-based method for detecting and recognizing characters in folktales. They concentrate on the interaction between a hand-crafted ontology of folktale characters and the linguistic analysis of indefinite and definite nominal phrases. Although promising, given the rather small scale of the experiment reported in this study, care must be taken when interpreting and generalizing the results. Using both syntactic features and a semantic lexicon, Goh *et al.* [5] propose an automated approach to the identification of the main protagonist in fictional texts, in particular fairytales. The method is not extended to the rest of actors.

Several studies have tried to tackle the problem of quote attribution. Quote attribution is the task of automatically extracting quoted speech and assigning them to their speakers. This task is interesting to us since we assume that the speakers of quoted speech and the conceptualizers of indirect speech form the cast of a story.

The studies about quote or speaker attribution can be divided into two types of approaches: rule-based approaches and machine-learning approaches. Rule-based approaches are represented by e.g. [4, 11, 3], who on the basis of syntactic rules and finite-state automata attempt to select the correct speaker of a piece of quoted speech. Other studies make use of machine-learning methods, e.g. [10, 3]. To find the correct speaker of a quote, Elson and McKeown [3] assign the quotes into one of seven syntactic categories. Many of these categories unambiguously lead to the identification of the speaker. For the remaining categories, a binary classifier is used that predicts for each pair of candidate speaker and quote whether it is the speaker or not on the basis of a feature vector. In contrast with most other work, Ruppenhofer *et al.* [11] incorporate not only direct speech, but also indirect speech and is thus more related to our work.

Most of the approaches mentioned above make use of external semantic lexicons to distinguish nominal phrases that are potential characters from other nominal phrases. The primary selection criterion is whether the nominal phrase falls within the category of living things. However, as stated above, whether some entity is animate or not is determined in relation to the story itself. The same lexicons are used to select sentences containing verbs of communication. In our approach we can do away with such external sources by formulating specific linguistic constructions. In doing so, our method is not restricted to, for example, verbs of communication, and can identify the more general category of verbs that describe some other mental state such as a thought, a belief or a desire. Another difference is that our approach is not limited to speaker identification of directly quoted speech, but incorporates the conceptualizer of indirect speech as well. Finally, our system is not restricted to the identification of an externally defined set of animate actors, but can in principle identify any subject of consciousness.

The method presented in this study not only tries to identify actors, but also attempts to rank them on a scale of importance. To our knowledge there are no computational studies that model the importance of actors in fictional texts.

3 Corpus, its annotation and tools

Corpus The corpus consists of a collection of 78 Dutch folktales from [12]. The collection is part of the Dutch Folktale Database from the Meertens Institute.[3] The collection was selected because it represents a homogeneous set of texts: (1) all texts are fairytales, (2) they are written in standard Dutch and (3) they are edited by the same author. The average number of words per story is 824 and a story consists of 44 sentences on average.

[3] `http://www.verhalenbank.nl`

Annotation To obtain gold-standard annotations of which actors are present in the texts, we asked two experts[4] in folktale research from the Meertens Institute to give a list of the actors of the stories. In addition to this list they were asked to rank the actors on a scale of importance to the story. Actors that were thought to be of equal importance should be placed in the same position in the ranked list. If a particular actor in a story is referred to with multiple names, we asked the experts to list all alternative names together in single entries in the ranking. We did not attempt to use or develop automatic procedures to resolve these coreferences.

Preprocessing and tools All texts were processed using the Dutch morpho-syntactic analyzer and dependency parser *Frog* developed at the ILK Research Group [14].[5] Frog contains a tokenizer that offers a high-precision quote detection system which was crucial for extracting all quoted sentences from the stories.

4 Methodology

The system consists of two procedures: one for direct and one for indirect speech. We developed a simple pattern matching algorithm for extracting the subject of sentences with direct speech. First we extract all sentences containing quoted speech (leaving out sentences that consist solely of quoted speech). Using the output of the chunking module of Frog, the system then searches for the nearest head of a noun phrase that is part of one of the following patterns:

(5) `QUOTE VP NP` (*"You, our king", cried the birds angrily.*)

(6) `VP NP QUOTE` (*"Oh dear", said the queen, "only look at my combs!"*)[6]

(7) `NP VP QUOTE` (*The fox asked, "How many tricks do you know?"*)

For indirect speech we extract the subject of matrix clauses of complementation constructions. We first locate all instances of subordinate conjunctions and interrogatives that have a verb complement dependency relation. We then extract the subject of the complement of the verb. We only included subjects that constitute a definite or indefinite noun phrase. Some examples are:

(8) *The queen wondered whether she would ever see her husband again.*

(9) *That the dragon was closing in, was obvious to the soldier.*

By applying both procedures to all texts we extract a set of actors for each text.

The set of actors should be ranked according to the actors' importance to the story. Regular term weight measures such as TF×IDF are not appropriate for this

[4] Theo Meder and Marianne van Zuijlen

[5] `http://ilk.uvt.nl/frog/`

[6] Note that this example also matches with the first rule. Unlike in English, Dutch uses inversion in subordinate clauses in which case the rule makes more sense.

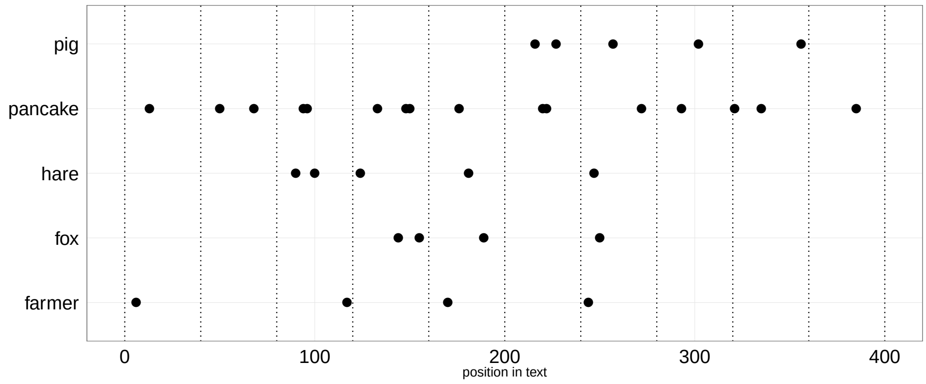

Figure 1: Lexical dispersion plot of actors in a version of *The Fleeing Pancake*.

ranking, because they estimate the importance of a term relative to the other texts in the corpus, whereas we would like to have some intra-textual measure of importance. At first sight, frequency of occurrence does not seem to be a good measure of importance either, because some actors might be referred to with high frequency in only a short text span and may not be of overall importance to the story.

We therefore propose a ranking method that is based on the dispersion of actors over a story. The basic idea is that more important actors are expected to appear at more places in the story and are more evenly distributed over the story than less important actors. Figure 1 plots the occurrences of the five actors in a version of *The Fleeing Pancake*. The pancake is present throughout the entire text whereas the fox, for example, is only briefly visible. Hence, we expect that the fox plays a less prominent role than the pancake.

We chose to use Juilland *et al.*'s [8] D statistical coefficient, because it has been reported as the most reliable dispersion measure [9, 190-191]. Juilland's D is calculated as:

$$D = 1 - \frac{V}{\sqrt{n-1}} \tag{10}$$

where n is the number of equally sized chunks in a text and V is the variation coefficient given by:[7]

$$V = \frac{\sigma}{\bar{v}} \tag{11}$$

where $\bar{v}$ is the mean frequency of a word in the different chunks and σ is the standard deviation of the frequencies in the chunks given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \bar{v})^2}{n-1}} \tag{12}$$

[7] The parameter n has to be set manually. In our experiments we observed little effect of the parameter within the interval $[5, 50]$ and set its value to $n = 20$.

rank	ground truth	subject baseline	our system
1	(Janneman, Jan, ventje 'little man')	Janneman	Janneman
2	(heks 'witch', wijf 'hag')	zak 'bag'	heks 'witch'
3	(tuinman 'gardener', haagknipper 'hedge cutter'), (arbeider 'worker'), (slootgraver 'ditch digger', man)	heks 'witch'	kat 'cat'
4	(kat 'cat')	huisje 'small house'	slootgraver 'ditch digger'
5	-	winkel 'store'	wijf 'hag'
6	-	pond 'pound'	haagknipper 'hedge cutter'
7	-	deur 'door'	arbeider 'worker'
8	-	kat 'cat'	-
9	-	neen 'no'	-

Table 1: Expert ranking and the results of the subject baseline (only the first 9 out of 33 results are shown) and our system for the story *Janneman in het papieren huisje* ('The Devil (Witch) Carries the Hero Home in a Sack'). Alternative names in the ground truth that refer to the same entity are placed between brackets.

Its values range from 0 (most uneven distribution) to 1 (most even distribution of a word across all chunks). To give an impression of some of the values different words can take, consider once more Figure 1. The vertical dotted lines mark the chunk boundaries at $n = 10$. Both the *farmer* and the *fox* are referred to four times, but because the farmer is more evenly distributed over the text it has a higher dispersion value, $D = 0.53$ and $D = 0.38$, respectively.

5 Experiment

For all documents in our corpus we extract a ranked list of actors. In the ideal case, the top of these lists contain the actors of a story. The extent to which this is the case reflects how well relevant items are found by our method and how well they are ranked according to the dispersion coefficient D. The results are evaluated by means of *Mean Average Precision* (MAP).

The subject of a sentence is often an entity that performs an action. Hence, subjects typically have a high chance of being an actor. We therefore compare our approach to a baseline in which all nouns and proper names are tagged as actors if they are classified by the dependency parser of Frog to hold a subject relation.

As an example, Table 1 presents the expert ranking, the results of the subject baseline (only the first 9 out of 33 results are shown) and the system for the story *Janneman in het papieren huisje* ('The Devil (Witch) Carries the Hero Home in a Sack'). Alternative names in the ground truth that refer to the same entity are

	subject baseline		system	
	complete	top three	complete	top three
frequency ranking	.739	.854	.909	.895
dispersion ranking	.792	.878	.918	.9

Table 2: MAP for both the subject baseline and the system using the frequency count ranking method and the dispersion based method. Scores are given for the complete result lists and for the top three items in the result lists against the first three ranks in the annotations.

		subject baseline		system	
		dispersion	frequency	dispersion	frequency
subject baseline	dispersion	–	$p < .0001$	$p < .0001$	$p < .0001$
	frequency		–	$p < .0001$	$p < .0001$
system	dispersion			–	$p < .03$
	frequency				–

Table 3: Pairwise Wilcoxon Signed-Rank Test using the dispersion and frequency ranking method for both the subject baseline model and our system.

placed between brackets. Since the annotations make no distinction between a canonical name of an actor and variants of that name, all alternative names are included in the set of relevant names. As noted in Section 3, we did not attempt to resolve coreferences. We thus make the simplifying assumption that all names in the result lists are to be considered as unique entities. The *Average Precision* (AP) for the baseline is .83. The system outperforms the baseline with a perfect score of 1.0, because it has identified solely relevant items.

We compare the proposed ranking method of actors to a method in which retrieved items are ranked according to their frequency of occurrence. The hypothesis is that the dispersion method should do a better job in distinguishing important from less important actors, because they are more important throughout the story and not only in small text spans.

Table 2 gives a summary of the results. The subject baseline model performs fairly well with a MAP of .739 using frequency ranking and .792 using dispersion ranking. Our system outperforms the baseline markedly with a score of .918. For our system, there is a minor difference between the two ranking methods.

To test whether the performance differences between the models are significant, we computed for all four models the AP for each document in the corpus. We then performed pairwise two-sided Wilcoxon Signed-Rank tests of which the resulting p values can be found in Table 3. All performance differences are sig-

		subject baseline		system	
		dispersion	frequency	dispersion	frequency
subject baseline	dispersion	–	$p > .2$	$p > .1$	$p > .2$
	frequency		–	$p > .05$	$p > .1$
system	dispersion			–	–
	frequency				–

Table 4: Pairwise Wilcoxon Signed-Rank Test using the dispersion and frequency ranking method for both the subject baseline and our system. AP is computed for the top three items in the result lists against the top three ranks in the annotations.

nificant. The significant differences between the two ranking methods support the hypothesis that dispersion is a helpful means to distinguish important actors from more peripheral actors.

Although the MAP shows that the system is well capable of retrieving mostly relevant items in the top retrieved results, it does not tell us whether the first, say, three items are the most important actors in a story. Recall that the annotations of the stories include a ranked list of actors. This allows us to evaluate whether the top retrieved results are not only relevant but contain the most important actors – as conceived by the annotators – as well.

The cast of a typical story contains only a few main actors and a range of supporting actors. It has one or two heroes, an opponent of the hero and either a wanted object or a helper of the hero. We recalculate the MAP scores using a constrained set of possible relevant actors in which only the first three ranks in the annotations are included. Using this set of relevant items, we compute the precision for the three highest ranked actors in the result lists (see Table 2). Again the system outperforms the baseline, albeit marginally, with a MAP of .895 using frequency ranking and .9 using the dispersion method and .878 for the baseline model using dispersion ranking and .854 using frequency ranking.

We performed the same Wilcoxon Signed-Rank tests as before to test for significance. Table 4 presents the results. We see that there are some small effects, but none of the performance differences are significant. This is interesting, because for the complete result lists, we saw significant differences between all combinations of systems (see Table 3). This suggests that the performance differences concentrate on the ranking of less important actors. It is also in this scope that the dispersion of a term is an effective indicator if its importance.

6 Conclusion

The approach taken in this paper to extract the cast from fictional texts proves to be successful. In line with literary scholars our results support the idea that intentionality serves as a strong feature to identify actors. The intentionality is reflected in the use of two related but structurally different linguistic constructions: direct and indirect speech. By extracting these constructions from a text we obtain a high-precision method for retrieving the cast of a story. The system performs significantly better than the baseline model that marks all nouns in subject positions as cast members.

Besides the mere identification of actors, we proposed a ranking method of the actors based on the dispersion of actors over a text. The hypothesis was that actors that are more evenly distributed throughout the text are more important than actors that only appear in short text spans. The results showed that ranking by means of dispersion gives a significant performance gain over 'simple' frequency counts.

We focus our recommendations for future research on two points. First, the system proposed in this paper depends heavily on the presence of direct and indirect speech. However, neither is necessarily present in every text. To make the system more robust, it should be able to deal with these texts as well. A good starting point to improve the stability is to use the output of the present system to learn about contexts other than direct and indirect speech in which actors appear.

We made the simplifying assumption that all retrieved items are unique entities, whereas in fact they corefer to a small subset of entities. Our second point of future work is directed towards resolving these coreferences, which will require adaptation of existing coreference resolution models to the domain of fictional texts.

7 Acknowledgments

The work on which this paper is based has been supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, as part of the Tunes & Tales project.

References

[1] Mieke Bal. *Narratology, Introduction to the Theory of Narrative*. University of Toronto Press, Toronto Buffalo London, third edition, 2009.

[2] Thierry Declerck, Nikolina Koleva, and Hans-Ulrich Krieger. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, 2012.

[3] David Elson and Kathleen McKeown. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1013–1019, 2010.

[4] Kevin Glass and Shaun Bangay. A naïve, salience-based method for speaker identification in fiction books. In *Proceedings of the 18th Annual Symposium*

of the Pattern Recognition Association of South Africa (PRASA '07), pages 1–6, 2007.

[5] Hui-Ngo Goh, Lay-Ki Soon, and Su-Cheng Haw. Automatic identification of protagonist in fairy tales using verb. In Pang-Ning Tan, Sanjay Chawla, Chin Ho, and James Bailey, editors, *Advances in Knowledge Discovery and Data Mining*, volume 7302 of *Lecture Notes in Computer Science*, pages 395–406. Springer Berlin / Heidelberg, 2012.

[6] Algirdas Julien Greimas. *Sémantique structurale*. Larousse, Paris, 1966.

[7] Walter Haeseryn. *Algemene Nederlandse Spraakkunst*. Groningen/Deurne: Marinus Nijhoff/Wolters Plantyn, 1997.

[8] Alphonse G. Juilland, Dorothy R. Brodin, and Catherine Davidovitch. *Frequency Dictionary of French Words*. Mouton de Gruyter, 1970.

[9] Michael P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh, 1998.

[10] Tim O'Keefe, Silvia Pareti, James Curran, Irena Koprinska, and Matthew Honnibal. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, 2012.

[11] Josef Ruppenhofer, Caroline Sporleder, and Fabian Shirokov. Speaker attribution in cabinet protocols. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, pages 2510–2515. European Language Resources Association (ELRA), 2010.

[12] Jacques Rudolf Willem Sinninghe. *Volkssprookjes uit Nederland en Vlaanderen*. Kruseman, Den Haag, 1978.

[13] Karina Van Dalen-Oskam. Names in novels: an experiment in computational stylistics. *Literary and Linguistic Computing*, 2012.

[14] Antal Van den Bosch, Bertjan Busser, Walter Daelemans, and Sander Canisius. An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. Van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium, 2007.

[15] Arie Verhagen. *Constructions of Intersubjectivity, Discourse, Syntax and Cognition*. Oxford Linguistics. Oxford University Press, Oxford New York, 2005.