Names in Novels: an Experiment in Computational Stylistics

Karina van Dalen-Oskam

Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences)
P.O.-box 90754
NL-2509 LT The Hague, The Netherlands

karina.van.dalen@huygens.knaw.nl

Abstract - Proper names in literary texts have different functions. The most important one in real life, identification, is only one of these. Some others are to make the fiction more 'real' or to present ideas about a character by using a name with certain meanings or associations to manipulate the reader's expectations. A description of the functions of a certain name in a certain text becomes relevant when the researcher can point out how it compares to the functions of other names and names in other texts. The paper describes how research into names in literary texts needs a quantitative approach to reach a higher level of relevancy. To get a first impression of what may be normal in literary texts, a corpus of 22 Dutch and 22 English novels and 10 translations into the other language in both sets was gathered. The occurrences of all names in these novels have been tagged for those data categories that seemed useful for the literary stylistic research planned. Some first results of the statistics are presented and the use of the approach is illustrated by means of an analysis of the use of geographical names in the Dutch novel *Boven is het stil* by Gerbrand Bakker and its English translation by David Colmer, *The Twin*. In the evaluation of the results, special attention is paid to the status of currently available digital tools for named entity recognition and classification, followed by a wish-list for the tools that this kind of research really needs.

1. Introduction

Names are so normal that many people overlook them. But they are important: we *are* our names, we take immense care in selecting names for our children, and most of us are not pleased when our names are misspelled or mispronounced. But also in other ways names are important. By name-dropping, using names of persons, places, artworks, buildings etc., we can show off what we know and who we know. So there is much more to names than 'just' identification. This applies even stronger to literary texts. Where we, in naming our children, can never make sure that they will 'live up' to the name we choose for them, the author of fiction does have that power.

Most readers and scholars will somehow have realized that some authors use names in a more subtle or more playful way than others. But until now, research in literary onomastics (the study of names in literary texts) is mostly qualitative by nature and often focuses on 'significant' names. No quantitatively comparative studies have been published yet. Several researchers, however, have pointed out that names can only be called significant if they are studied in the context of all the names - the so-called 'onymic landscape' - in a text, oeuvre, genre etc. (e.g. Sobanski 1998). This question is comparative by nature and implies the wish for a more quantitative, and thus computer-assisted approach. It is expected that different authors, genres, time periods or even languages apply different name types and name functions in different ways, showing different trends which we want to discover in what we like to call comparative literary onomastics (Van Dalen-Oskam 2005, 2006).

2. Name types and name functions

My aim is to analyze the stylistic functions of name usage in literary texts, wanting to be able to compare the usage and functions of names across texts, oeuvres, genres, time periods, and cultures or languages. For that, a quantitative approach seems useful. As to name usage, I want to know how many name forms usually occur in a novel, and how many of these are personal names, place names, and other names. This looks like an easy task: just counting names according to the types they belong to. Name functions are certainly less easy to quantify. The main function of all names is to identify a person, place, or object - to distinguish a unique person, place, etc. from other persons, places, etc. However, in literary texts several other functions can play a role. A very important function of names usually is to make the fiction more 'real' (called the 'reality-enhancing' function) (Debus 2002: 76-77). Furthermore, an author can subtly present ideas about a character or place etc. by using a name with certain meanings or associations, which are thus used to manipulate the reader's expectations as to the personality of characters ('Characterization', or when less direct 'Association') (Debus 2002: 77-81). And ofcourse a name can have more than one function.

As to name functions, another useful disctinction is between *plot internal* and *plot external* names. Plot internal names refer to characters, places or other entities which only ´exist´ in the fiction of the story. Plot external names refer to persons, places or other

entities which are known to exist or to have existed in the real world. Most place names in novels are plot external, referring to real countries, cities, streets, etc. and thus have a reality enhancing function. In fantasy novels, however, place names are usually fictional and thus plot internal, enhancing the unreality, the fantasy of the story. Plot external personal names often seem to function as characterizations of the fictional characters, describing e.g. their political or cultural preferences.

The amount of names in a text can be expressed in the percentage of the total amount of tokens in the text. For that, we need digital texts of fair to good quality. Different forms of the same name (e.g. *Patrick* and *Patrick's*) need to be grouped by a ***lemma*** (*PATRICK* in this case). Different name forms for the same person or place need to be related to the same ***entity*** (e.g. the name *Alfred* and the name *Issendorf* both belong to the character identified as *ALFRED ISSENDORF*). To find out whether we can compare the resulting percentages across languages, we focused on a corpus of modern Dutch and English novels and their translations into the other language.

We found we had to include two other levels of aggregation: ***mentions*** and ***name tokens***. Mentions are occurrences of a name irrespective of the number of tokens used. So several name tokens can be used in one mention. This distinction is necessary because different languages have different tokenization rules. The Dutch personal name *Gerrit-Jan*, e.g., with a hyphen and therefore counted as one token, is translated in English as *Gerrit Jan*, resulting in two tokens.

Comparative research can only be done when many scholars collaborate. We will have to make sure that all those scholars encode their texts in the same way, considering the same tokens as names. This may sound easy, but it is not. Even name theorists have different definitions of what a name is (Van Langendonck 2007). Guidelines had to be set. We decided to limit the tagging to the 'prototypical' names, so those names that are considered names by readers in general. Something is a name if it refers to a unique person, place, or object. So we excluded currencies, days of the week, months, etc. For cases leading to discussion we defined additional rules, which we will not go into here.

The name categories taken into account in the tagging were divided over three main categories: personal names, geographical names, and names of other things. Personal names were subdivided in first names, family names, and nicknames. First names and family names have further subdivisions (Dutch corpus only) which will not play a role in the analyses addressed in this paper. The interesting division of names into plot internal and plot external will also not be very prominent in this paper, which will focus on the general picture for all names and the main name categories of personal names and geographical names.

3. The corpus

Currently not many modern novels are available in digital form yet, which meant we had to digitize most of the corpus ourselves. Due to intellectual property rights we cannot make these digital files available to other scholars. A lot of manual work combined with the use of perl-scripts was needed (more about this is Section 6.3) before the first analyses could be done, which is why the corpus is still relatively small. It consists of 22 Dutch novels and 22 English novels, added with the translation into the other language of

10 in each group. 4 of the 22 novels in each language are books for children/young adults. Altogether the corpus consists of 64 novels and 4.499,999 tokens.
The Dutch language corpus contains 20 novels from The Netherlands and 2 from Flanders. The English language corpus contains 13 novels from the United States, 5 from Great Britain, and 4 from Canada. The focus of the corpus is on the last 20 years (cf. Table 1). The novels that were selected are listed in Table 4.

Years	Dutch	English transl.	English	Dutch transl.
40's-50's	1		1	
60's	3	1	2	2
70's	2		3	
80's	2	1	2	2
90's	5	2	5	1
10's	9	6	9	5
Total:	**22**	**+ 10 +**	**22**	**+ 10 = 64**

Table 1 Diachronic distribution of selected novels (authors and titles are listed in Table 4)

4. First impressions of the onymic landscape

When we look at the absolute amounts of name tokens compared to the total amount of tokens in the novels (Fig. 1) we see that the trend is that the longer a novel is, the more name tokens it contains. This may not be very surprising. The picture partly shifts, however, when we draw up the graph replacing the number of name tokens on the horizontal axis by the number of different named entities (only available for the Dutch corpus).

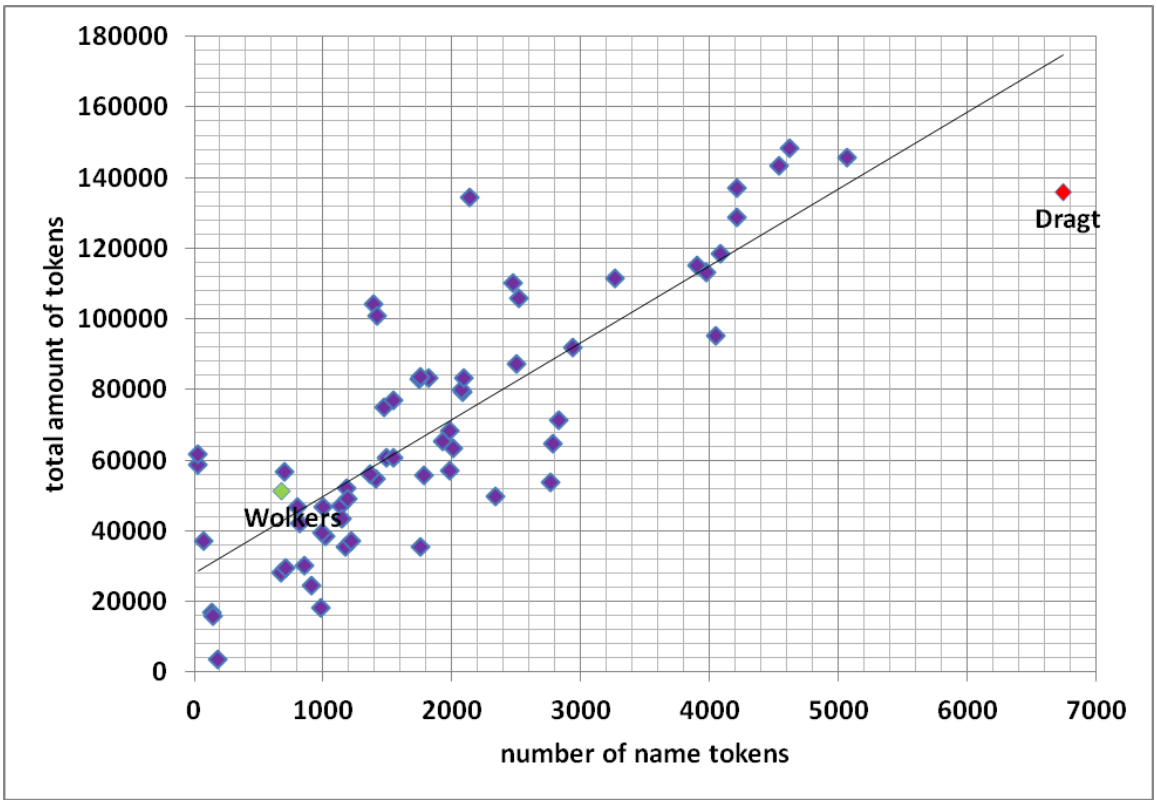

Fig. 1 The 44 original novels and the amount of name tokens (horizontal axis) related to the total amount of tokens (vertical axis). Examples marked: the novel by Dragt and the novel by Wolkers.

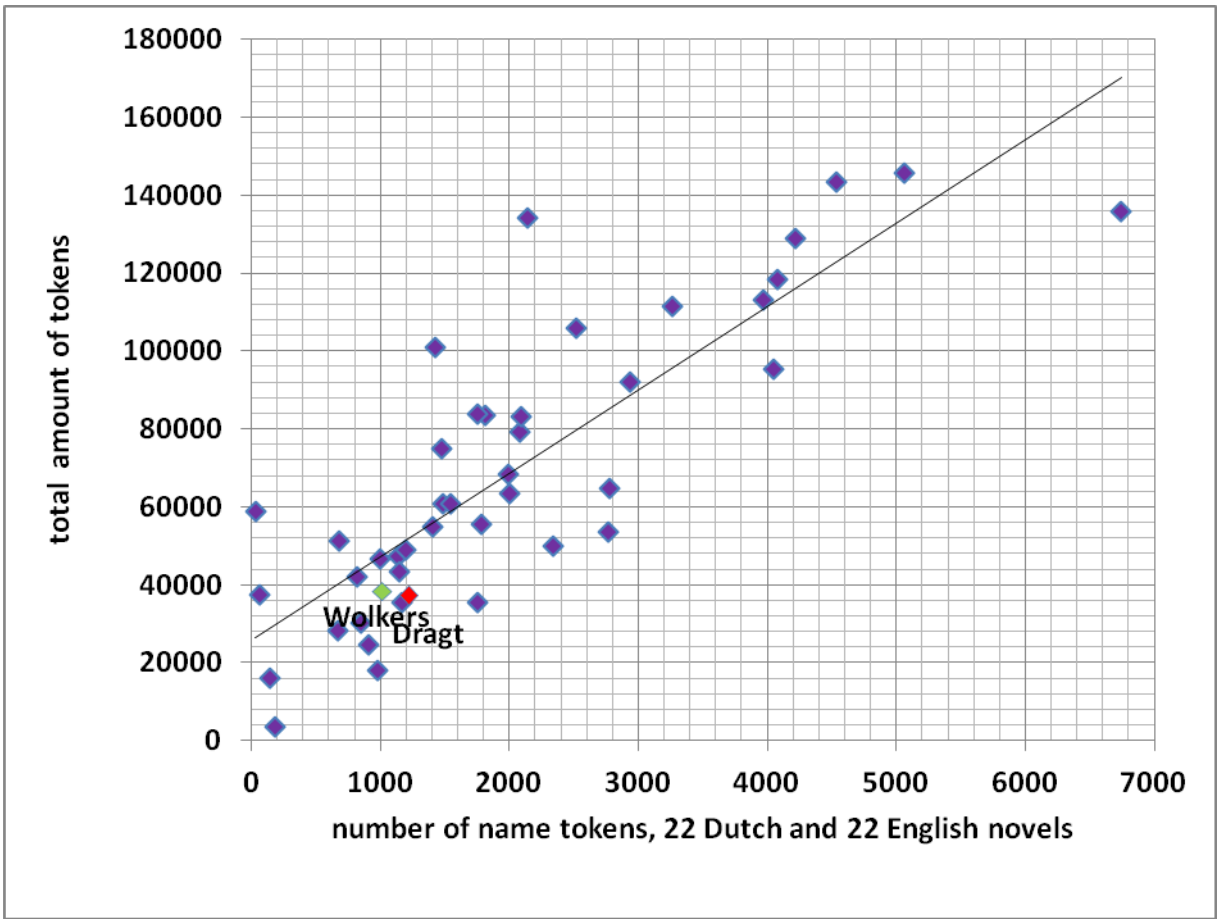

Fig. 2 The 22 original Dutch novels and the amount of entities which have a name (horizontal axis) related to the total amount of tokens (vertical axis). Examples marked: the novel by Dragt and the novel by Wolkers.

One of the two example novels from the Dutch corpus highlighted in Fig. 1 shows up on a totally different place in the graph in Fig. 2: the Dragt book does have an unusually high amount of name tokens, but only a very small amount of named entities, clearly breaking the trend in Fig. 2. This shows that we should be careful in drawing conclusions based on only one quantitative approach to the data.

It is useful, however, to present an overview of the percentages of the amount of name tokens of the total amount of tokens in all of the 44 different novels. Fig. 3 shows these, sorted from the highest percentage at the top to the lowest at the bottom of the bar chart. Most of the novels have a name token percentage of about 2 % to about 3.5 %.

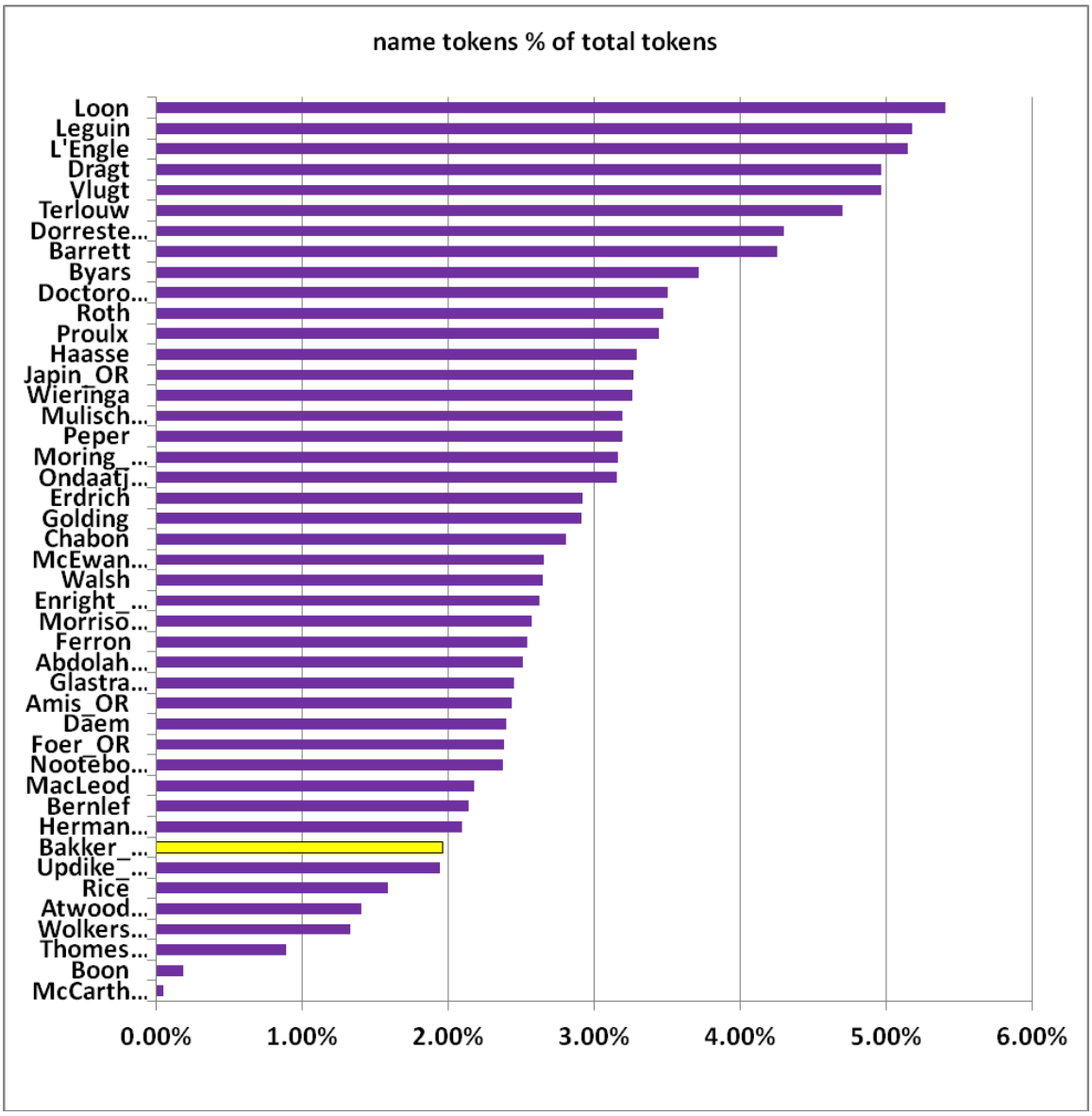

Fig. 3 Percentage of name tokens of the total amount of tokens; Dutch and English novels are nicely mixed in the ranking

What is interesting in Fig. 3 is that the six novels at the top of the bar chart, with the highest percentages, are all books for children or young adults (Loon, LeGuin, L'Engle, Dragt, Vlugt, Terlouw). The two others are on place 9 (Byars) and - much farther down - place 24 (Walsh). The probable cause for the high percentage of name tokens in these books for children / young adults is that names are often used where other texts would have a pronoun. We will not extend our research to verify this hypothesis, however.

For this contribution, we will next focus on the results in the originals and translations in the corpus. Several elements other than the percentages of name tokens need to be taken into account for a comparison of name usage in different languages: differences in the total amount of tokens in the original and the translation, and differences in the amount of mentions used in the original and the translation.

Author (year OR, TR)	% OR	% TR	OR > TR	Mentions	Diff mentions
Updike (1966, 1967)	1.95	1.72	1.11	784 > 775	−9
Amis (1991, 1991)	2.44	2.26	1.07	1073 > 1056	−17
McCarthy (2006, 2010)	0.05	0.05	1.05	28 > 28	0
Foer (2002, 2002)	2.38	2.25	1.04	2067 > 2067	0
Atwood (1985, 1987)	1.41	1.33	1.03	1197 > 1187	−10
McEwan (2007, 2007)	2.66	2.50	1.03	850 > 845	−5
Ondaatje (1987, 1989)	3.16	2.95	1.03	1598 > 1588	−10
Morrison (1970, 1984)	2.58	2.43	1.03	1324 > 1300	−24
Doctorow (2004, 2006)	3.51	3.38	1.02	3689 > 3709	+20
Enright (2007, 2007)	2.63	2.59	1.01	1943 > 1950	+7

Table 2 Original English novels and their Dutch translations. The columns show the author and the year of publication of the original and the translation, the percentage of name tokens in the original (% OR) and in the translation (% TR), the change in total amount of tokens between original and translation in percentages (OR > TR), the change in total amount of mentions between original and translation (Mentions) and the difference in mentions between original and translation (Diff mentions).

Table 2 shows the relevant data for the ten English novels and their translation into Dutch. As to the difference in total amount of tokens in original and translation we can see that all translations had more tokens than the original and that only the oldest of the selected novels (Updike) was enlarged with more than 10 % (cf. Rybicki 2010). The percentage of name tokens was also slightly higher in all translations, except in the novel by McCarthy, which has an extremely low amount of name tokens anyhow in which the translator did not change anything. Comparing the amount of name mentions in originals and translations shows that only two tranlations show more mentions than the original (7 and 20 more in total), two showed no difference, and six have less mentions in the translation, ranging from 5 to 24.

Author (year OR, TR)	% OR	% TR	OR > TR	Mentions	Diff mentions
Wolkers (1963, 1967)	1.33	1.23	1.11	638 > 640	+2
Dorrestein (1996, 2003)	4.30	3.96	1.10	2560 > 2589	+29
Thomése (2003, 2005)	0.89	0.80	1.07	109 > 106	−3
Japin (1997, 2000)	3.27	3.07	1.06	3428 > 3397	−31
Abdolah (2000, 2006)	2.51	2.86	1.05	1945 > 2069	+124
Nooteboom (1991, 1994)	2.38	2.41	1.04	594 > 621	+27
Möring (1997, 1999)	3.16	3.11	1.03	4199 > 4255	+56
Bakker (2006, 2009)	1.97	2.00	1.03	1388 > 1386	−2
Mulisch (1983, 1985)	3.20	3.48	1.02	1614 > 1745	+131
Hermans (1966, 2006)	2.10	2.11	0.99	1642 > 1608	−34

Table 3 Original Dutch novels and their English translations. The columns show the author and the year of publication of the original and the translation, the percentage of name tokens in the original (% OR) and in the translation (% TR), the change in total amount of tokens between original and translation in percentages (OR > TR), the change in total amount of mentions between original and translation (Mentions) and the difference in mentions between original and translation (Diff mentions).

Table 3 shows the data for the ten analyzed Dutch originals and their English translations. The results are different in several places compared to Table 2. As to the total amount of tokens, nine out of the ten novels are enlarged by the translator. One was reduced, but only to 99 % of the original. Two novels were enlarged with ten percent or more, one of them also the oldest in the subcorpus (Wolkers). The resulting name token percentages are now a bit more, now a bit less in the translations than in the originals. The big difference with Table 2 is to be found in the results for the name mentions, which are clearly opposite to those of the Dutch translations from the English originals: Four novels have less mentions in the translation, namely 2, 3, 31, and 34. The other six, however, have more, one with only 2 more, but the other five having many more mentions more than in Table 2, ranging from 27 to 131. From both Table 2 and Table 3 we gather that in this very small corpus, the translators from Dutch to English seem to have had a rather different approach to name usage than the translators from English to Dutch. Further research into this, e.g. on a much larger corpus, will be needed to find out whether language differences play a role here, or if for instance there are cultural differences which lead English translators from Dutch to explication by means of names more often than their counterparts, translating English into Dutch.

5. Case: The Twin

To illustrate how the analysis of the usage of names can lead to new insights in a literary work and into possible functions of names We will present one case, namely a Dutch novel written by Gerbrand Bakker, *Boven is het stil* (first published in 2006), and its English translation *The Twin* (translated by David Colmer and published in 2009). In 2010, the novel won the prestigious International IMPAC Dublin Literary Award.

When we asked some Dutch colleagues and friends whether they thought this novel contains a lot of names, their reply was that it would have hardly any names. An American reader of the English translation, however, came up with a totally different reply. He stated that the novel seemed to have a lot of names, especially geographical names. This anecdote, however simple, is a useful starting point for a quantitative analysis. We have different readers with totally different intuitions about the same stylistic element. Who is right?

Let's have a look at the most general numbers. In Fig. 3, we find that the Bakker novel (in the original Dutch) has a percentage of name tokens of 1.97 %, which is at the lower end of what seems to be normal for the corpus as we analyzed it. The intuition of the American reader therefore seems to be wrong, but this does not mean that the intuition of the Dutch readers is proved right - the percentage is clearly not to be described as "hardly any names", as the Dutch readers expected it to be. Can it be that the

English translation has a lot more names than the original? Table 3 shows that this is not the case. The English translation hardly shows any differences as to the name tokens and the name mentions. So influence of the translation on the intuition of the American reader can be ruled out. But what about the other statement of this reader, that the novel has a lot of geographical names?

Sigla	Author *Translator*	Title (year)	PERS	GEO
Abdolah_OR	Kader Abdolah	Spijkerschrift (2000)	1630	376
Abdolah_TR	*Susan Massotty*	My father's notebook (2006)	1854	545
Amis_OR	Martin Amis	Time's arrow (1991)	818	257
Amis_TR	*Gideon den Tex*	De pijl van de tijd (1991)	801	258
Atwood_OR	Margaret Atwood	The handmaid's tale (1985)	974	114
Atwood_TR	*Gerrit de Blaauw*	Het verhaal van de dienstmaagd (1987)	983	120
Bakker_OR	Gerbrand Bakker	Boven is het stil (2006)	1096	314
Bakker_TR	*David Colmer*	The twin (2009)	1113	357
Barrett	Andrea Barrett	The forms of water (1993)	3276	496
Bernlef	J. Bernlef	Hersenschimmen (1984)	632	143
Boon	Louis Paul Boon	Menuet (1948)	49	11
Byars	Betsy Byars	Bingo Brown, Gypsy Lover (1990)	792	14
Chabon	Michael Chabon	The final solution (2004)	552	171
Daem	Geertrui Daem	Koud (2001)	867	165
Doctorow_OR	E. L. Doctorow	The march (2004)	2782	941
Doctorow_TR	*Sjaak Commandeur*	De mars (2006)	3089	653
Dorrestein_OR	Renate Dorrestein	Verborgen gebreken (1996)	2261	186
Dorrestein_TR	*Hester Velmans*	Crying shame (2003)	2298	193
Dragt	Tonke Dragt	De brief voor de koning (1962)	5851	606
Enright_OR	Anne Enright	The gathering (2007)	1638	292
Enright_TR	*Piet Verhagen*	De samenkomst (2007)	1649	292
Erdrich	Louise Erdrich	The plague of doves (2008)	2873	240
Ferron	Louis Ferron	Het stierenoffer (1975)	1349	106
Foer_OR	Jonathan Safran Foer	Everything is illuminated (2002)	1551	522
Foer_TR	*Peter Abelsen*	Alles is verlicht (2002)	1562	504
Glastra van Loon	Karel Glastra van Loon	Lisa's adem (2001)	1227	112
Golding	William Golding	The inheritors (1955)	1987	0
Haasse	Hella S. Haasse	Fenrir (2000)	818	176
Hermans_OR	Willem Frederik Hermans	Nooit meer slapen (1966)	1220	386
Hermans_TR	*Ina Rilke*	Beyond sleep (2006)	1187	398
Japin_OR	Arthur Japin	De zwarte met het witte hart (1997)	2836	944
Japin_TR	*Ina Rilke*	The two hearts of Kwasi Boacha (2000)	2792	975
LeGuin	Ursula LeGuin	Jane on her own (2003)	171	0

L'Engle	Madeleine L'Engle	A wrinkle in time (1962)	2571	89
Loon	Paul van Loon	Dolfje weerwolfje (1996)	954	1
MacLeod	Alistair MacLeod	No great mischief (1999)	761	730
McCarthy_OR	Cormac McCarthy	The road (2006)	18	9
McCarthy_TR	*Guido Goluke*	De weg (2010)	18	9
McEwan_OR	Ian McEwan	On Chesil Beach (2007)	560	280
McEwan_TR	*Rien Verhoef*	Aan Chesil Beach (2007)	571	265
Moring_OR	Marcel Möring	In Babylon (1997)	3558	622
Moring_TR	*Stacey Knecht*	In Babylon (1999)	3599	626
Morrison_OR	Toni Morrison	The bluest eye (1970)	1225	128
Morrison_TR	*Nettie Vink*	Het blauwste oog (1984)	1186	126
Mulisch_OR	Harry Mulisch	De aanslag (1983)	1333	248
Mulisch_TR	*Claire Nicolas White*	The assault (1985)	1470	258
Nooteboom_OR	Cees Nooteboom	Het volgende verhaal (1991)	352	212
Nooteboom_TR	*Ina Rilke*	The following story (1994)	369	220
Ondaatje_OR	Michael Ondaatje	In the skin of a lion (1987)	1150	421
Ondaatje_TR	*Graa Boomsma*	In de huid van een leeuw (1989)	1148	408
Peper	Rascha Peper	Vingers van marsepein (2008)	2207	317
Proulx	E. Annie Proulx	Postcards (1992)	2746	866
Rice	Anne Rice	Interview with the vampire (1976)	1464	558
Roth	Philip Roth	The human stain (2000)	3079	1005
Terlouw	Jan Terlouw	Oorlogswinter (1972)	1963	278
Thomese_OR	P. F. Thomése	Schaduwkind (2003)	62	36
Thomese_TR	*Sam Garrett*	Shadow child (2005)	61	39
Updike_OR	John Updike	Of the farm (1966)	673	111
Updike_TR	*Frans Bijlsma*	Terug en verder (1967)	679	92
Vlugt	Simone van der Vlugt	Schuld (2007)	1675	28
Walsh	Jill Paton Walsh	A chance child (1978)	1058	40
Wieringa	Tommy Wieringa	Alles over Tristan (2002)	908	134
Wolkers_OR	Jan Wolkers	Een roos van vlees (1963)	544	70
Wolkers_TR	*John Scott*	A rose of flesh (1967)	549	76

Table 4: amounts of name tokens for personal names (PERS) and geographical names (GEO) in the 44 original novels and in the translations

As to name tokens in the novels in the corpus, geographical names occur much less than personal names (cf. Table 4). Acting characters in a text seem to need a lot more explicit mentioning than the places where they are or where they travel to. But another view on the name landscape is to have a look at the amount of *different* names in these two categories of personal names and place names - the amount of lemmas. In Fig. 4, the amount of different personal names and different geographical names in each of the 44 original novels is visualized in a bar chart. Fig. 4 shows that most of the novels have a clearly higher amount of different personal names than geographical names and that a much smaller set have more different geographical names - and that the Bakker novel is

the most extreme in this. This seems to prove the intuition of the American reader right on this point and leads us to have a closer look at the geographical names in the novel: what are they used for and how do they function?

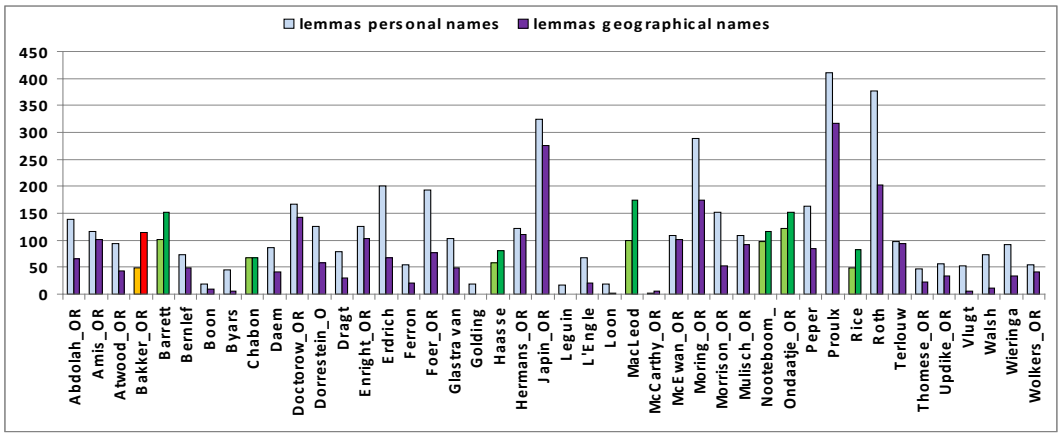

Fig. 4 For each of the 44 original novels the left bar shows the amount of different personal names (lemmas) and the right bar show the amount of different geographical names (lemmas). Eight novels show more different geographical names than personal names, the biggest difference being found in the Bakker novel *Boven is het stil*

The main character of *The Twin* is Helmer van Wonderen, a small farmer in the area to the North of Amsterdam, who at the age of fifty-five starts to free himself from the life that was forced upon him. When he studied Dutch Language and Literary Studies at the University of Amsterdam for seven months, his twin brother Henk died in a car crash. Henk was the farmer of the two and destined to take over the farm. After Henk's death, their father put Helmer in Henk's role. Thirty years later, when the health of Helmer's father is slowly deteriorating, Helmer finally starts to deal with Henk's death and with shaping his own life into something he can be happy with.

On closer inspection, the novel shows two significant usages of geographical names. The first of these is the fact that when Helmer is going somewhere, to a funeral or to do some shopping, we repeatedly read the exact route he takes by car. We will present only one example here:

> In Monnickendam I take the N247 and follow it to Edam, where I drive through the village to the dyke, because if I don't get off here I'll be stuck on the main road to Oosthuizen. Near Warder I stop the car for a moment to have a better look at a flock of birds. (42)

These excursions mostly go to the area to the North of Amsterdam. Helmer himself seems to explain this in a remark about his thoughts on the day his twin brother died:

> Since that day almost every journey I make is north. I no longer go south of the village. (60)

A second peculiarity is connected to Helmer giving shape to his own life. At the beginning of the novel he puts his bedridden father upstairs, redecorating the ground floor bedroom for himself, buying a big bed and modern linnen. Then his neighbour Ada suggests that the room needs 'some art' to make it complete. Helmer then buys an old geographical map of Denmark, has it framed, and puts it on his bedroom wall. Repeatedly he walks to the map and reads lists of names aloud. One example:

> I close my bedroom door and go over to stand in front of the map of Denmark. 'Helsingør, ' I say. 'Stenstrup, Esrum, Blistrup, Tisvildeleje.' Five names spoken slowly are not enough tonight. I do a few extra islands: 'Samsø, Aerø, Anholt, Møn. ' The big bed is ready for me. (238)

It becomes clear that reading aloud lists of Danish names from the map functions as a mantra for Helmer, to help him get an emotional grip on things. Apart from this special function of the Danish geographical names, the explicit travel routes to the North of Amsterdam emphasize the territorial taboo of the main character ("I no longer go south") - which is a second special function. Both stylistic functions of names in this novel were not apparent before the quantitative analysis on the corpus of novels described above.

6. Evaluation

6.1 Evaluation of the research results

Many more observations can be made about the first measurements of names in the pilot corpus, of which only a few could be addressed in this paper. The normal range of name token percentage in novels seems to be lying between a bit below 2 % to around 3.5 %. Children's books seem different: seven of the eight analyzed books for children and young adults score above 3.5 %.

Another interesting result for the very small corpus of originals and translations is that English translators from Dutch seem to add more name mentions than Dutch translators from English. This needs to be tested on a bigger corpus before we can proceed to look for an explanation.

We think the most interesting result of our pilot is that the quantitative approach not only helped to verify and falsify readers' intuitions, but that a closer look at the ways in which these intuitions differed and comparing them to the statistical outcome leads to a new research question about the processing of names by readers. The working hypothesis based on the anecdotal case here presented would be that a reader who knows the geographical names he reads tends to overlook them, whereas to the reader who does not know them their unfamiliarity makes them stand out. Thus, it would mean that the general knowledge or the educational level of the reader will have a very concrete influence on how the geographical names in the novel are processed. Knowing the names leads to underestimating their role, and not knowing them leads to an overestimation. Seen from this perspective, the two functions of geographical names we identified for the novel *Boven is het stil / The Twin* are more subtly cloaked in the Dutch original than in the English translation. This difference is again closely related to the fact that the translation is a faithful one.

6.2 Evaluation of the method

We experienced that it is sometimes difficult for a scholar to encode novels from a culture that is not her own. An example is that a Dutch scholar sometimes does not know whether a personal name in an American novel is a first name, a family name, or a nickname. The same will go the other way round. This means that a comparative approach of the analysis of name usage and name functions can only be reliably done when scholars from different cultures join their efforts. We tested this on a very small scale, by comparing the name tags added to two American novels by an American scholar with what a Dutch scholar would do. It proved that the choice to only encode prototypical names and drawing up a set of guidelines for difficult cases was quite workable and did only very occasionally lead to different choices. To make sure that inter- and intra-encoder-agreement is acceptable this needs to be tested, however, on a larger corpus to be enriched by a larger group of scholars.

6.3 Evaluation of available data and tools

We found that not many modern novels are digitally available yet and if so, they were not (or not exhaustively) tagged for names. This meant a lot of work, which meant the pilot corpus could not be as large as we would have wished. The hope that existing named entity recognition and classification tools (NERCs) would speed up the tagging of the names proved entirely false. Since for many digital humanists this seems to be a quite unexpected evaluation, we will go into this in a bit more detail.

A good overview of the state of NERC can be found in (Sekine and Ranchod 2009). We tested several NERCs on one novel in the English corpus. Based on this first test we selected the Stanford NER (the 4 class version) as the best and decided to apply this to the other English-language novels in the corpus. After doing this for three more novels, we decided this took too much time and reverted to a kind of manual tagging (more about this later).

The Stanford 4 class NER yielded a very satisfactory result in that it recognized about all the tokens we wanted to tag as names and more. This meant we had to only remove tags we did not agree with. But on further analyzing the classes the NER added to our files, we found that what especially makes the NER good in another context made it quite annoying when we applied it to one novel. For each and every name occurrence, the tool decides which class to attribute it to. And this meant that the same name for the same person, place, or object, could (and did) get different classes assigned to it at different occurrences. This meant that at each occurrence of the same name the attributed class tag needed to be checked. Since when analyzing one novel at a time we are usually pretty sure that all occurrences of the same name belong to the same peron, place, or object, this was extremely frustrating. It proved to be easier and most of all *faster* to tag the ASCII file by searching a name and replacing it by the name surrounded by the correct tag.

What also needs to be emphasized is that the NERC tools only address a part of the tags we wanted to add. The NERC tools do not have subcategorizations such as we wanted, e.g. for personal names to subdivide them into first names, nicknames, and family names. And on top of that, the distinction between plot internal and plot external names could not be tagged using the existing tools. This was a further exhortation to use the search-an-replace-approach described above, which could include the

subcategorization and the plot internal versus external tagging all in one and the same iteration.

The last issue we want to mention is that to calculate the percentages of name tokens in a novel we had to calculate the total amount of tokens in that novel as well. We found that tokenization is done differently in different tools, and again is looked upon differently in different language areas. We found for example that several tokenization calculations for English counted words such as "didn't" and "I'm" as one token. For Dutch scholars it is normal to see these words as enclitic forms and to count them as two tokens. It is important to take these differences into account, since using different tokenization approaches on the same corpus of novels could lead to differences in the statistics which have nothing to do with actual differences in the texts. For our corpus of Dutch and Englih texts we therefore used a very simple token-counting perl script that was applied to all of the novels in the corpus, independent of the language.

What we would need to speed up the type of analysis described in this paper is a virtual research environment in which the scholar can upload the digital text, and can add tags to the text in different ways. One way could be to run a NER that provides only the basic tag "name" to a token, and the suggestions of which the scholar can either approve or reject. Then in a next phase, using a concordance option and filtering and sorting options, the scholar could add subcategorizations and the tag for plot internal or external names. Another way could be to have an advanced search and replace option, which helps the scholar to select from a concordance which of all occurrences of a token should be provided with what kind of name tags. Apart from these tagging options, we would need easy tools to view the statistics for all added tags, related to the total amount of tokens per text. In sum: we need an environment which integrates all kind of tools which could be applied to the texts in different phases of the research, building on the idea that a scholar needs to be able to add knowledge to her corpus all the time and to visualize that knowledge on the fly whenever this helps the exploration of the data.

7. Conclusion

In this paper, we could only show a small part of all interesting observations to be made about the usage of names in a corpus of modern Dutch and English novels. But these first results make us anxious for more, in the expectation that this approach may lead to an acceptable method for a.o. across language comparison of stylistic elements. We conclude that the preliminary results are sufficiently interesting to go into the stylistic analysis of name usage and functions in novels more deeply. Names also seem promising stylistic elements for a comparison across languages. The currently available tools which could be expected to be helpful for this type of research, proved to be insufficient. We therefore plan to develop a set of interrelated webservices which will assist the scholar in the recognition, categorization, further tagging, and statistical analysis of names in novels.

Acknowledgements

Many thanks to David L. Hoover, whose help was invaluable in making the whole procedure of tagging a lot more efficient. The Excelsheet with macros he designed for this purpose is a great help in visulaizing what kind of tools are needed for this kind of

analysis. Our discussions about the stylistic functions of names and related topics in computational stylistics were also very inspiring.
As always, André van Dalen was extremely helpful in writing perl scripts for any kind of labour intensive task that presented itself during the research.

References

Debus, F. (2002). *Namen in literarischen Werken. (Er-)Findung - Form - Funktion*. Stuttgart: Franz Steiner Verlag, Akademie der Wissenschaften und der Literatur, Abhandlungen der geistes- und sozialwissenschaftlichen Klasse, Jahrgang 2002, nr. 2

Rybicki, J. (2010). Original, Translation, Inflation. Are All Translations Longer than Their Originals? Poster presented at DH2010, London, 7 - 10 July 2010, http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-841.pdf

Sekine, S. and Ranchod, E. (2009). *Named Entities. Recognition, classification and use.* Amsterdam/Philadelphia: John Benjamins.

Sobanski, I. (1998). 'The Onymic Landscape of G. K. Chesterton's Detective Stories.' In: *Proceedings of the XIXth International Congress of Onomastic Sciences, Aberdeen 1996*. 3 Vols. Aberdeen. Vol. 3, p. 373-378

Stanford NER (2009)**.** Stanford Named Entity Recognizer (NER) version 1.1.1. http://nlp.stanford.edu/software/CRF-NER.shtml (accessed 6 October 2011)

Van Dalen-Oskam, K. (2005). 'Vergleichende literarische Onomastik'. In: Brendler, A. und S. Brendler (Hrsg.). *Namenforschung morgen: Ideen, Perspektiven, Visionen.* Hamburg: Baar, 2005, p. 183-191. English translation, 'Comparative Literary Onomastics', at http://www.huygens.knaw.nl/wp-content/bestanden/pdf_vandalenoskam_2005_Comparative_Literary_Onomastics.pdf

Van Dalen-Oskam, K. (2006). 'Mapping the Onymic Landscape'. In: Maria Giovanna Arcamone, Donatella Bremer, Davide de Camilli e Bruno Porcelli (eds.): *Il nome nel testo. Rivista internazionale di onomastica letteraria* VIII (2006), p. 93-103. Atti del XXII Congresso Internatzionale di Scienze Onomastiche, Pisa, 28 agosto – 4 settembre 2005 (*Proceedings of the 22nd International Congress of Onomastic Sciences, Pisa 28. Aug. – 3. Sept. 2005)*, vol. III.

Van Langendonck, W. (2007). *Theory and typology of proper names*. Berlin / New York: Mouton de Gruyter.