Namescape:
or How to Deal with Noise

Karina van Dalen-Oskam
Jesse de Does
The Netherlands

Abstract

Researchers of literary onomastics have increasingly called for the analysis of all the names in a literary work, an oeuvre, a genre or a time period. This approach, however, is humanly impossible. But things are changing. Digital text corpora and the tools to recognise, classify and tag named entities are becoming more and more available, along with tools to search, retrieve and visualise the data. The results of the automatic analysis of a large corpus of text, however, still include a lot of 'noise' – mistakes made by the software. Nevertheless, the number of correct results can be staggering. To leave them aside because of the noise would not be wise, but that is still the reaction of many humanities scholars. We propose a different reaction, namely to try to find ways to deal with noise and thus profit from the large gains that are there for the taking. This paper presents some first methodological experiments in dealing with noise that were carried out during the Namescape project (www.namescape.nl), which concerned modern Dutch fiction.

* * *

Introduction

The aim of comparative literary onomastics is to analyse the usage and functions of proper names in literary texts such as novels, plays and poems (van Dalen-Oskam 2005, 2006). The approach is based on the hypothesis that patterns and trends can be discovered in the ways in which literary authors make use of proper names in their work. These patterns involve the number of names used as well as the distribution of and visible preferences for different functions in which names are applied. Some examples of name functions are identifying or (its opposite) cloaking an identity, and characterising the entity endowed with the name. For both usage and function, a quantitative analysis is considered to be appropriate. The patterns in usage and functions are expected to be related to either genre or, for example, topics or thematic characteristics in the works on the one hand, and to the background and preferences of the authors on the other hand. The onymic choices an author makes are assumed to be consciously or unconsciously informed by her or his cultural background and knowledge, and are therefore expected to differ between authors and texts. They may also differ throughout the ages. As such, comparative literary onomastics approaches the use of proper names in literary works as embedded in the broader, literary and general culture of a time period and area.

To verify or falsify these fundamental hypotheses we need to have access to a huge amount of data on name usage and name functions in literary works. This implies a very large corpus of literary and other kinds of texts (to be able to compare between literary and other text types) – 'very large' being something like many thousands, preferably millions of texts,

Names and Their Environment. Proceedings of the 25th International Congress of Onomastic Sciences, Glasgow, 25-29 August 2014. Vol. 5. *Literary Onomastics. Other Names. Commercial Names*. Carole Hough and Daria Izdebska (eds)
First published 2016 by University of Glasgow under Creative Commons licence (CC BY-NC-ND 4.0)

nicely metadated with information about author, date of publication, text type and so on.[1] This corpus needs to have all the proper names tagged, with additional labelling of the name type and the name functions. To give only one example: many occurrences of a proper name in a literary work will have the function of identifying an entity (a character), but may also characterise the entity. Since it is impossible to tag a really large corpus manually, we need to turn to information technology to see whether some of these steps can be automated.

Automatically detecting proper names in texts and classifying the name occurrences as belonging to different types of names has a long history under the label 'named entity recognition and classification' (NERC or NER). These tools were originally developed for non-literary texts. How useful the application of NER tools to literary texts would be from the perspective of comparative literary onomastics was one of the research questions in the Dutch project Namescape. The project ran from April 2012 until December 2013 and was funded by CLARIN-NL, the Dutch branch of the European-wide digital infrastructure project CLARIN (Common LAnguage Research Infrastructure), which is intended for humanities researchers who work with language data and tools. In this paper, we report some of the main results of Namescape.

Namescape

The full name of the project was *Namescape: Mapping the Onymic Landscape*. It was a 'demonstrator project', that is, a project scheduled to deliver a set of tools (computer programs) geared towards literary onomastic research. The collaborating partners were the Huygens Institute for the History of the Netherlands (Huygens ING – KNAW, one of the research institutes of the Royal Netherlands Academy of Arts and Sciences), the Institute for Dutch Lexicology (INL) and the Information and Language Processing Systems Group of the University of Amsterdam (ILPS – UvA). A large corpus of books in digital form (around 8,000 titles) was gathered, but due to copyright issues the files of most of these cannot be made publicly available. The corpus was annotated with a rich tag set. The proper names in the whole corpus can be searched, and search results are presented as keywords in context; in other words, the names are presented in concordance form, with a couple of words before and after the keyword so that the context of the name can be gleaned from this snippet of text. Furthermore, several visualisation tools help the scholar explore the huge amount of data that has thus become available.

The Namescape project took as point of departure a pilot project in which the usage of proper names in a corpus of 22 Dutch and 22 English novels was analysed (van Dalen-Oskam 2013). As to name functions, the pilot project focused on the use of geographical names in these texts (van Dalen-Oskam 2012, 2013). That usage and functions of place names are especially interesting has also been convincingly demonstrated by Rosa and Volker Kohlheim (Kohlheim and Kohlheim 2011, Kohlheim 2013). The tagging of this small corpus was done using a combination of manual and semi-automatic procedures. The tagging took

[1] Opinions differ on how large such collections have to be. From a humanities perspective, I rather like the definition that Alan Riddell presents in a recent article: 'I will refer to any collection of texts as a *very large collection* if it contains more texts than a single researcher would be expected to digest in a year's worth of dedicated reading' (Riddell 2014: 92).

around 12 months. The tagset was specifically developed for the research questions from the perspective of comparative literary onomastics, as sketched above. We wanted to be able to make separate calculations for the occurrences of first names and family names, because our hypothesis is that the usage of first names has functions that differ from the usage of family names. For instance, the level of 'intimacy' in a novel could be indicated by the ratio between first and family names (van Dalen-Oskam 2005). We also wanted to distinguish names referring to entities (characters, places, etc.) that 'exist' only within the fictional narrative from names referring to persons and places that exist in the real world. We label the first ones 'plot internal' (e.g. Bilbo, Tyrion) and the second ones 'plot external' (e.g. Merkel, Glasgow). The reason for this distinction is the hypothesis that names that refer to plot-internal entities usually have functions that differ from those that refer to plot-external entities. Only the systematic labelling of name occurrences for this feature will enable the testing of this hypothesis.

Although many tools for automatic named entity recognition and classification (NER, NERC) exist, these two distinctions are generally not made in the tools. In the Namescape project, we not only tested existing NER tools, but also retrained the tool that best suited our needs and developed a new one for Dutch novels. We presented several visualisation tools to explore the results. More information and links to the tools can be found at www.namescape.nl. In this contribution, we focus on the fact that NER tools are still far from perfect, suggesting possible ways to deal with the 'noise' – the mistakes – in the results of such tools.

Named Entity Recognition

In the semi-automatically and manually corrected pilot corpus of 44 novels, personal names are far more frequent than place names, and occurrences of other names are so sparse that they are statistically uninteresting. This is different for the text types on which NER tools are usually trained, such as newspapers. As can be expected, existing NER tools are especially elaborate in their distinction of different classes of other names, such as organisations, brands, etc. As stated, existing NER tools lack a subdivision for personal names with the types 'first name', 'family name' and 'nickname', and they do not label names as being plot internal or plot external. And when an entity is mentioned by first name and family name, traditional NERs tag the complete phrase as a name, whereas from a literary onomastic point of view we would want to consider these as two different name types that in combination indeed do indicate a certain entity (character), but of which the first name can have a set of functions that is different from the set of functions of the family name. To be able to answer onomastic questions about literary names, we therefore need another kind of NER.

The best available NER proved to be the Stanford NER (van Dalen-Oskam *et al*. 2014), which was developed primarily for English-language texts. When we applied the Stanford tagger to Dutch literary texts, training on Dutch-language newspapers, the results were below expectations. This only partly had to do with the language difference. We improved the results for Dutch literary texts by using a special training corpus. The adapted Stanford tagger was the first new NER we delivered. The second one – the 'Namescape

tagger' – added the classification of personal names into first name / family name / nickname and expanded the structure of the tags dealing with combinations of first names and family names, giving each part of the name a separate tag and clustering the parts together as a whole.

We evaluated the success rate of the two new NERs. The overall score (F1) for the adapted tagger was 0.845; for the Namescape tagger, it was slightly better, namely 0.893. If we look at the scores for the different name types, personal names and place names have the highest F scores. The tagger adapted from the Stanford NER scored 0.896 for personal names and 0.776 for place names. Again, the Namescape tagger performed slightly better, namely 0.936 for personal names and 0.844 for place names. Both taggers scored badly on other names, such as organisations: 0.339 with the adapted Stanford NER and 0.222 (Miscellaneous) and 0.54 (Organisation) with the Namescape tagger. Since in the pilot project with 44 novels this name category proved to be not very frequent, meaning that a statistical approach would not be very relevant, this low score is certainly not a big problem.

Although the scores for the successful recognition of personal names and place names seem rather high, there are still a lot of errors. And these are what the human eye tends to notice more than all the correct attributions. Names are missed by the tool, words that are certainly not names are marked as names, and names are miscategorised.

Named Entity Resolution

We have not yet been able to adapt the taggers to have them label names as referring to plot internal or plot external entities. In the Namescape project, we applied an existing tool for named entity resolution that we hope will help us find a way to do this. We applied a 'semanticiser', a tool used for something called 'Wikification'. The tool makes a list of words or phrases (N-grams) in a running text that may be represented with an entry in Wikipedia, and provides each item on this list with a direct link to the Wikipedia entry that most probably describes the item, based on probability scores, based on internal link probabilities within Wikipedia, that are calculated for each possible link (van Dalen-Oskam *et al.* 2014).

In the pilot study we found that most names in the chosen 44 novels were plot internal. This group included, for example, the names of the characters, with the names of the main characters occurring most often. Names referring to plot external entities often occur only once or twice in a text, and tend to be either real place names or the names of well-known politicians, artists and so forth. For now, we assume that names that are found by the Wikifier and linked to a Wikipedia entry will most probably refer to plot external entities. We know this is not a perfect match: J.R.R. Tolkien's famous hobbit character Bilbo Baggins has his own Wikipedia entry, as does the relative newcomer Tyrion Lannister, from George R.R. Martin's series of fantasy novels *A Song of Ice and Fire*, better known under the title of the first volume, *A Game of Thrones*.[2] But these are expected to be outliers. Furthermore, the tool attempts to detect articles about fictional entities using features extracted from the entry title and the categories that are attributed to the entry.

[2] Bilbo Baggins, see Wikipedia (2014a) s.v. 'Bilbo Baggins', and for Tyrion Lannister see Wikipedia (2014b) s.v. 'Tyrion Lannister'.

We have not yet been able to measure the success rate of the Wikifier. It seems clear that further adaptation for the application to the literary domain is desirable, because when we looked at the list of generated links, we immediately noticed a lot of errors. For one novel, for instance,[3] the name 'Stalin' got the correct link to the entry in the Dutch Wikipedia, but strangely enough 'Adolf Hitler' did not get a link.[4] The names of several fictional characters did not get a link, as was expected; however, others did get a link to entries presenting real persons with the same name. And the name of one of the female main characters ('Jet') also got a link, but to the Wikipedia entry describing an Australian rock band of the same name. This is typical: the most prominent type of error is perhaps over-resolution: isolated first name or surname parts referring to a plot internal character are often resolved to an apparently unconnected Wikipedia entry. One has to take into account here that the current version of the semanticiser has been designed to optimize the choice between different possible resolutions, rather than the decision between resolution and non-resolution. Despite all this, however, reasonable success rates in classifying entities as plot internal or external have been measured on the pilot corpus.

How to Deal with Noise?

The reaction of many scholars who see the errors thrown up by new digital tools such as the ones described above, is to decide never to use the tools and to disregard any other deliverables produced in the era of digital humanities. However, NER tools do not perform any worse than other tools; on the contrary, they tend to be among the most successful ones. Errors – 'noise' – are a standard feature of many digital tools. The question is: how useful are tools that still make this many mistakes?

The tools can deal with a lot more texts and a lot more data than an individual scholar or even a group of scholars could ever deal with, let alone in a reasonable amount of time. On top of that, the number of correct hits is overwhelmingly larger than the number of misses. We therefore believe that it would be unwise to ignore these new tools: we will simply have to learn how to deal with noise.

We will be able to cope with any errors by applying two strategies. The first entails quantifying the uncertainty caused by noise by comparing manually tagged with automatically analysed data. How many names were found by the taggers, and how many were tagged as such by the scholar, manually or using semi-automated procedures? What is the ratio? This ratio can then be used to estimate the reliability of results on data that have not been tagged manually. In the Namescape project, we did this to see whether the results of the Wikifier tool were somehow comparable to the results of the manual tagging for names referring to plot internal and plot external entities (van Dalen-Oskam *et al*. 2014). The results of this measurement were not perfect, but we established a rather high correlation between the manual and the automatic ratio (Pearson r = 0.87). So even though the automatic tagging resulted in quite a lot of mistakes (from the perspective of the human eye), and it is

[3] The novel is *De tweeling* by Tessa de Loo. See http://visualizer.namescape.nl/book/73 (accessed 11 December 2014).

[4] On closer inspection we found this has to do with a systematic error in the reading of the name through OCR (optical character recognition).

unrealistic to assume that all names with links to Wikipedia refer to plot external entities, we can use the number of established links as an indicator or a predictor of the number of names that refer to plot external entities in a text or a corpus.

The second strategy follows from the first. We are talking about predictions and indications, and certainly not about definitive proof or convincing results leading to the confirmation of a hypothesis or the falsification of an assumption. Many humanities scholars automatically assume that computer tools are meant to be perfect, and that the tools' makers meant them to deliver absolute truths leading to clear verification. This is an unrealistic expectation. It is more profitable to look at the tools as a way to explore much more data than ever before and to do so in much more depth, and thus find pointers for the next steps in the research. The results can, for instance, inspire the scholar to focus on certain parts of the dataset for a more detailed analysis to answer a specific research question. This means that we should not consider the tools the deliverers of absolute truth in the whole heuristic process of knowledge creation. For the literary onomastics scholar, the tools are not the goals but additional means (in addition to, for example, tools such as close reading) to arrive at new answers.

The visualisations that many tools generate to present their results also have to be seen as a means to explore the data. Here, we mention only one of the types of visualisations created in the Namescape project; more can be found through the Namescape website. Screenshots can be found in van Dalen-Oskam *et al*. (2014), where technical details of much of the above can also be found. In the barcode graph (see Figure 1), we can follow the occurrence of personal names throughout a novel. The visualisation does not give the absolute total number of occurrences of a name, but uses a bar to indicate for each paragraph of the file whether a name occurs in that paragraph. The paragraphs are automatically numbered and run from left to right on the horizontal axis. The visualisation helps scholars easily pinpoint which characters occur in which parts of a novel. Main characters are visible throughout the novel; secondary characters show up only now and then or in a specific part of the novel.

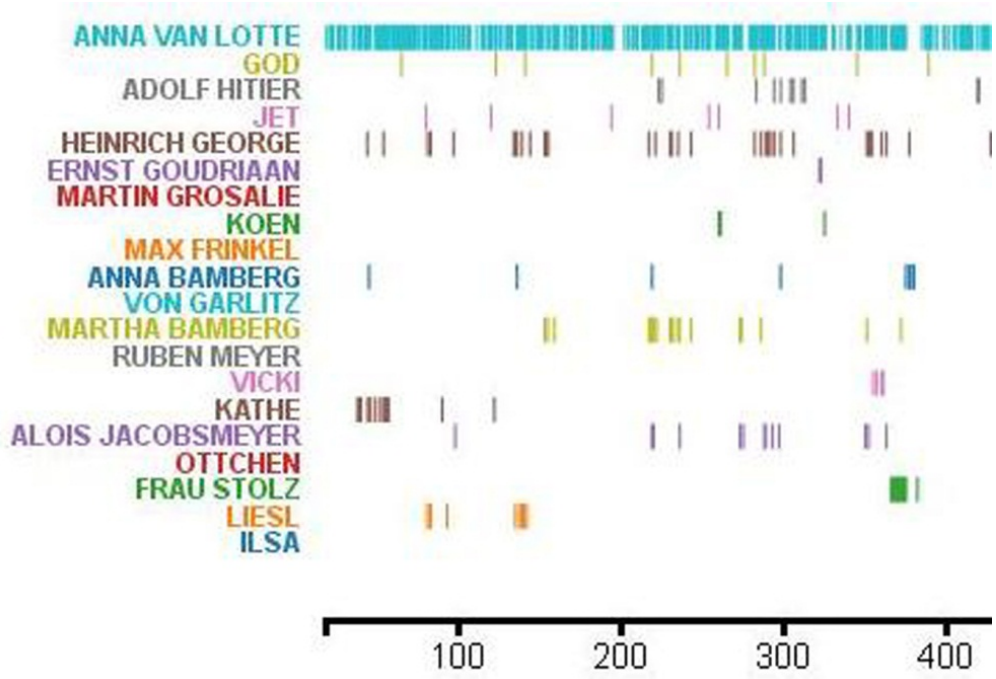

Fig. 1. The barcode graph of Tessa de Loo's *De tweeling*.
See http://visualizer.namescape.nl/graph/index/type/barcode/bookid/73 (Date of access: 11 December 2014). The occurrence of a name in a paragraph is indicated on the horizontal axis by means of a bar.

Even when we know that the underlying tagged file of the novel may have an error rate of, say, 10%, this should normally not distort the general pattern of (co)occurrence we see. The visualisation can help scholars to develop new ideas about the usage and functions of personal names in a novel, and can be the starting point for a more detailed analysis of the usage pattern of, for example, a small set of the personal names or a small part of the corpus or of one novel.

Conclusion

Although the tools are not yet perfect and the number of errors seems staggeringly high, there is no reason not to use them for research into literary onomastics. The hits already outweigh the misses, meaning that the amount of data we can explore is much larger than ever before. Taking into account that it may still be too early to use these results as certain proof, we will be able to benefit from them by using them as indicators of certain patterns and as such as useful means to better explore large amounts of data. The explorations, using every possibility that visualisation techniques have to offer, will certainly lead to new ideas, new hypotheses and new kinds of research. In addition to exploring the data and the results, literary onomastics scholars should sit down with information technology specialists and tool developers to devise ways in which the existing tools can be further adapted to support their own research. It is in these kinds of collaborations, where each of the participants contributes what she or he is best at, that far-reaching progress can be achieved.

Karina van Dalen-Oskam
Huygens Institute for the History of the Netherlands/
University of Amsterdam
The Netherlands
karina.van.dalen@huygens.knaw.nl

Jesse de Does
Institute for Dutch Lexicology
The Netherlands
jesse.dedoes@inl.nl

References

van Dalen-Oskam, K. (2005) 'Vergleichende literarische Onomastik'. In: Brendler, A. and Brendler, S. (eds.) *Namenforschung morgen: Ideen, Perspektiven, Visionen.* Hamburg: Baar. 183-191. English translation, 'Comparative Literary Onomastics'. Available online 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: Arcamone, M.G., Bremer, D., de Camilli, D. and Porcelli, B. (eds.) *Il nome nel testo. Rivista internazionale di onomastica letteraria VIII: Atti del XXII Congresso Internatzionale di Scienze Onomastiche, Pisa, 28 agosto – 4 settembre 2005*. Vol. 3. Pisa: dall'ETS. 93-103.

van Dalen-Oskam, K. (2012) 'Immer nach Norden. Gebrauch und Funktion von Eigennamen\ im Roman *Oben ist es still* von Gerbrand Bakker. Ein Pilotprojekt zur vergleichenden literarischen Onomastik'. *Beiträge zur Namenforschung Neue Folge* 47.1. 33-58.

van Dalen-Oskam, K. (2013) 'Names in Novels: An Experiment in Computational Stylistics'. *LLC: The Journal of Digital Scholarship in the Humanities* 28. 359-370.

van Dalen-Oskam, K. *et al*. (2014) 'Named Entity Recognition and Resolution for Literary Studies'. *Computational Linguistics in the Netherlands Journal* 4. 121-136. http://www.clinjournal.org/sites/default/files/09-VanDalenOskam-etal-CLIN2014.pdf

Kohlheim, R. and Kohlheim, V. (2011). 'Der literarische Name zur Jahrtausendwende: Andreas Maiers Roman *Wäldchestag* als Beispiel'. *Beiträge zur Namenforschung Neue Folge* 46.3. 269-285.

Kohlheim, V. (2013) 'Urbanonyme in der Literatur: Funktion und Status'. In: Kremer, D. and Kremer, D. (eds.) *Die Stadt und ihre Namen* 2. Vol. 2. Leipzig: Leipziger Universitätsverlag. 327-350.

Riddell, A.B. (2014) 'How to Read 22,198 Journal Articles: Studying the History of German Studies with Topic Models'. In: Erlin, M. and Tatlock, L. (eds.) *Distant Readings. Topologies of German Culture in the Long Nineteenth Century* Rochester, NY: Camden House. 91-113.

Wikipedia (2014a) ‘Bilbo Baggins’. Date of access: 15.12.2014. Available online at: http://en.wikipedia.org/wiki/Bilbo_Baggins

Wikipedia (2014b) ‘Tyrion Lannister’. Date of access: 15.12.2014. Available online at: http://en.wikipedia.org/wiki/Tyrion