CHAPTER 30

Namescape: Named Entity Recognition from a Literary Perspective

Jesse de Does[a], Katrien Depuydt[a], Karina van Dalen-Oskam[b,c] and Maarten Marx[c]

[a]Dutch Language Institute, Leiden, jesse.dedoes@ivdnt.org, katrien.depuydt@ivdnt.org;
[b]Huygens Institute for the History of the Netherlands, The Hague, karina.van.dalen@huygens.knaw.nl;
[c]Universiteit van Amsterdam, Amsterdam, maartenmarx@uva.nl

ABSTRACT

The project Namescape: Mapping the Landscape of Names in Modern Dutch Literature (2012–2013) was a demonstrator project granted in the third CLARIN-NL call. Partners in the project were the Huygens Institute for the History of the Netherlands, the University of Amsterdam, and the Dutch Language Institute (CLARIN centre). The project dealt with Named Entity Recognition (NER) for modern Dutch fiction and delivered two new NER tools for this purpose. It also addressed Named Entity Resolution and focused on a set of visualisations of names in individual texts from the corpus. This chapter gives an overview of the results of the project, starting with a description of the background of the research questions in the discipline of comparative literary onomastics. It then goes on to describe the tools that were delivered, and which can be found on the project website, http://www.namescape.nl/.

30.1 Introduction

The research discipline dealing with name studies – onomastics – includes a subdiscipline in which scholars aim to analyse and compare the usage and the function of names in literary works. In this kind of research, which can be called comparative literary onomastics, the scholar assumes that patterns and trends can be discovered in the way in which literary authors make use of proper names in their work (van Dalen-Oskam, 2005, van Dalen-Oskam, 2016). The comparative literary onomastics analysis not only deals with quantitative issues, such as the amount of names in a work, but also with a more qualitative evaluation of the functions of the names that have been

How to cite this book chapter:
de Does, J, Depuydt, K, van Dalen-Oskam, K and Marx, M. 2017. Namescape: Named Entity Recognition from a Literary Perspective. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 361–370. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.30.

used. Names almost always have an identifying function, to discriminate one place or person from another, but in some literary cases names are also used to do the opposite, that is to hide a person's identity or the location of a place. And, to give just a few examples of other functions, personal names may also be used to describe the personality of a certain character, and place names clearly help to situate a story in a specific geographical area or to emphasise that area as imaginary.

The problem that the Namescape project wanted to address is that this type of literary onomastics until now could deal with only one text or a very small corpus, due to a lack of specialised tools for named entity recognition and classification in literary texts. The most efficient approach available was to privately scan texts and manually annotate them, whereas the need clearly was to be able to compare name usage and name functions in much larger corpora of literary works. Direct incentive for the Namescape project was a pilot performed by literary onomastician van Dalen-Oskam on a collection of 22 Dutch and 22 English novels, in which the use of proper names was analysed (van Dalen-Oskam, 2013). Tagging this corpus, using a combination of semi-automatic and manual tagging, took around 12 months. The literary named entity recognition applied in the pilot differs from usual named entity recognition in two respects: (1) personal names, place names, and other names were tagged. Personal names were also tagged as being a first name, a family name, or a nickname. This was necessary from the perspective of literary onomastic analysis to be able to test the hypothesis that first names and family names may be used with different effects and different functions. Furthermore, the literary onomastician needs to view these as separate instances of separate names and not lumped together as one name. References to a character with only a first name, only a family name, or a combination of both may each have a different stylistic effect and a different (set of) function(s). (2) All names were further labelled with information on whether they were purely fictional, referring to 'plot internal' entities (e.g. Harry Potter), or referred to 'plot external', really existing, named entities (e.g. Churchill, London).[1] This was done to be able to test the hypothesis that 'plot internal names' and 'plot external names' have a different set of functions.

The conclusion of the pilot was that a much larger corpus of literary works was needed to confirm or correct the observations that were made in the pilot project, helping the scholar to perform statistically significant quantitative analysis of the use of names, and so was a set of tools for the researcher to tag, search and analyse the corpus, including insightful visualisations. The Namescape project set out to do just that.

30.2 Namescape Research Environment Components

The main tasks for the Namescape project were: to create a larger corpus, to perform good quality Named Entity (NE) recognition on literary material and try to perform NE resolution so as to determine whether names in literary works are plot internal or plot external, and finally to make the data available in an environment in which the researcher can search and visualise search results (technical details can be found in van Dalen-Oskam et al., 2014).

30.2.1 Backend: Corpora and Annotation Tools

30.2.1.1 Namescape Corpora and Annotation Scheme

For the core NE corpus, the project took the Dutch part of the corpus that van Dalen-Oskam used (the 'Huygens corpus') in the above mentioned pilot, consisting of 22 Dutch novels and containing ca 1.5 million tokens, and extended it with a collection of 550 OCRed Dutch books from the period 1970–2009 and containing ca 28 million tokens.

[1] When real, existing persons are acting characters in a novel, their names were tagged as plot-internal.

Random paragraphs were selected from that extended core corpus in order to create a manually annotated gold standard corpus for NE recognition, consisting of about 1 million tokens. There were two reasons to compose the gold standard corpus in this way. First, annotating a limited selection of complete works would have severely limited the amount and variety of name mentions in the training corpus.[2] Furthermore, by choosing snippets of annotated texts instead of complete texts as training material we hoped to circumvene IPR issues, so as to be able to distribute the training corpus for research purposes.[3] For evaluation of NER performance, we used a fixed random split of the corpus in a training and a testing partition.

In the course of the project, three additional corpora were collected and curated: a corpus of eBooks (over 7,000 books and ca 500 million tokens), a subselection of the SoNaR Corpus[4] (over 100 books and ca 11 million tokens) and a corpus with the Dutch books from the Gutenberg project (530 books and ca 30 million tokens). The last corpus contains books from the 17th to 20th century, which is a challenge for NE recognition because of the historical Dutch spelling.

The XML encoding was done in TEI P5. We made a simple extension to TEI (Text Encoding Initiative) to tag the named entity properties. We also chose to use a single tag for named entities and a different tag for entity parts to avoid nested name tags. For the basic principles for NE recognition, we have followed the 1999 Named Entity Recognition Task Definition Chinchor et al., 1999. The basic definitions (quoted from section 30.3 of the Task Definition) are:

> PERSON: named person, family, or certain designated non-human individuals
> ORGANIZATION: named corporate, governmental, or other organizational entity
> LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.) and astronomical locations (Chinchor et al., 1999).

We refer to the Task Definition document for further details.

We added a MISC category for name occurrences which did not clearly fit in any of the three basic types.

30.2.1.2 Named Entity Recognition

We have used our Namescape training corpus and trained two named entity taggers: the Namescape-trained instance of the Conditional Random Field-based Stanford tagger[5] (with the default settings), and a Support Vector Machine-based (SVM) tagger,[6] which has been designed to improve performance by making use of information derived in an unsupervised way from a corpus (in this case the extended core Namescape corpus). Both taggers are fairly standard supervised machine learning applications with slightly different, but similar, feature sets, consisting of a set of context features and a set of word shape features. The SVM tagger had a slightly better performance (cf. van Dalen-Oskam et al., 2014). Table 30.1 gives an evaluation of NER performance, using a fixed random split of the gold standard corpus. The application of NER to the other corpora has not been evaluated in a strict way; manual inspection shows a roughly similar accuracy on the

---

[2] Cf. also Landsbergen (2012), where the two approaches are compared, and better performance is shown for the snippet approach.

[3] We hope to make the training corpus available from http://www.inl.nl/taalmaterialen.

[4] Oostdijk et. al. (2008), Oostdijk et. al. (2013).

[5] The Stanford NER tagger is described in Finkel et al. (2005) and is available from http://wwwnlp.stanford.edu/ner/ We have used version 1.2.6, from 2012-07-09.

[6] The Namescape Support Vector Machine-based tagger has been developed at the INL using SVM^{light} (http://svmlight.joachims.org/) by way of the Java native interface JNI SVM-light-6.01 (http://adrem.ua.ac.be/~tmartin/).

Tagger	NE type	precision	recall	F1
Stanford	location	0.802	0.712	0.754
	misc	0	0	0
	organisation	0.433	0.228	0.299
	person	0.876	0.895	0.881
	overall	0.853	0.824	**0.838**
Namescape	location	0.83	0.729	0.776
	misc	0	0	0
	organisation	0.516	0.251	0.339
	person	0.867	0.917	0.896
	overall	0.853	0.838	**0.845**

Table 30.1: Tagging accuracies obtained on the Namescape gold standard corpus.

eBooks and SoNaR corpora, and, as was to be expected given the differences in language, a much worse accuracy on the historical Gutenberg data.

It should also be mentioned that, taking advantage of the annotation of the pilot corpus, we endeavoured to go beyond the standard NER annotation categories by distinguishing between first names, family names and nicknames, thus accommodating the wish of the literary name scholar to compare, for example, the usage and functions of first names with that of family names, instead of heaping them all together as personal names.

Web Service and Application Since we wanted to enable non-technical users to do named entity recognition on their own texts, we created a small lab environment which has both NE taggers implemented as a web service and is easy to use (http://ner.namescape.nl/namescape/tagger). Text can be uploaded in several formats (plain text, HTML, EPUB, Word, TEI) from the user's own computer or directly from the web by supplying a URL. The result of the tagging process is a TEI file with the inline annotation, delivered to the user either as is ('raw output') or formatted and displayed with NEs highlighted. The formatted display also includes overviews of names per category, snippets per name and a co-occurrence graph allowing the user to explore the relations between the named entity mentions.

30.2.1.3 Named Entity Resolution

To establish whether a name is plot internal or plot external, NE resolution has been performed by means of the ILPS semanticiser (Odijk et al., 2013; cf. http://semanticize.uva.nl/doc/). The tool tries to link named entities in the texts to entries in Wikipedia (a process also known as wikification). A name is considered to be plot-external when the entry in Wikipedia describes a non-fictitious entity.[7]

The application of the method does not require a manually annotated training corpus. For evaluation, the pilot corpus was used, in which plot-internality and plot-externality is manually tagged. For the 3862 distinct name types and 35852 name tokens, we have obtained a type accuracy of 74.5% and a token accuracy of 79.8%. The most prominent type of error is perhaps over-resolution: plot-internal entity mentions are often resolved to an apparently unconnected Wikipedia entry. This is understandable when we take into account that most proper names in a novel are expected to be plot-internal, and that the semanticiser has been designed to optimise the choice between

[7] The fictitiousness features we used were whether the article title or category contained any variant of 'legend', 'mythological' or 'fictional'.

different possible resolutions, rather than the decision between resolution and non-resolution (for more details see van Dalen-Oskam et al., 2014).

30.2.2 Front End: Search Interface and Visualisations

30.2.2.1 Search Interface

To enable the user to search and browse the texts, a search interface was built using XQuery on an eXist XML database (http://search.namescape.nl; see Figure 30.1). Unfortunately, IPR restrictions forced us to limit access to the full texts for part of the corpus.

30.2.2.2 Visualisations

Visualisations of NEs in a single text are enabled through the Namescape visualiser. Visualisation of NEs in a corpus is done via the barcode browser.

Namescape Visualiser The Namescape Visualizer[8] (http://visualizer.namescape.nl/) gives an overview of the names in a text and shows the co-occurrence of names in paragraphs. To create a picture of the onymic landscape of proper names in the novel *De vergaderzaal* by A. Alberts, you may select the novel in the tool, and then an overview is given of the top twenty most frequent

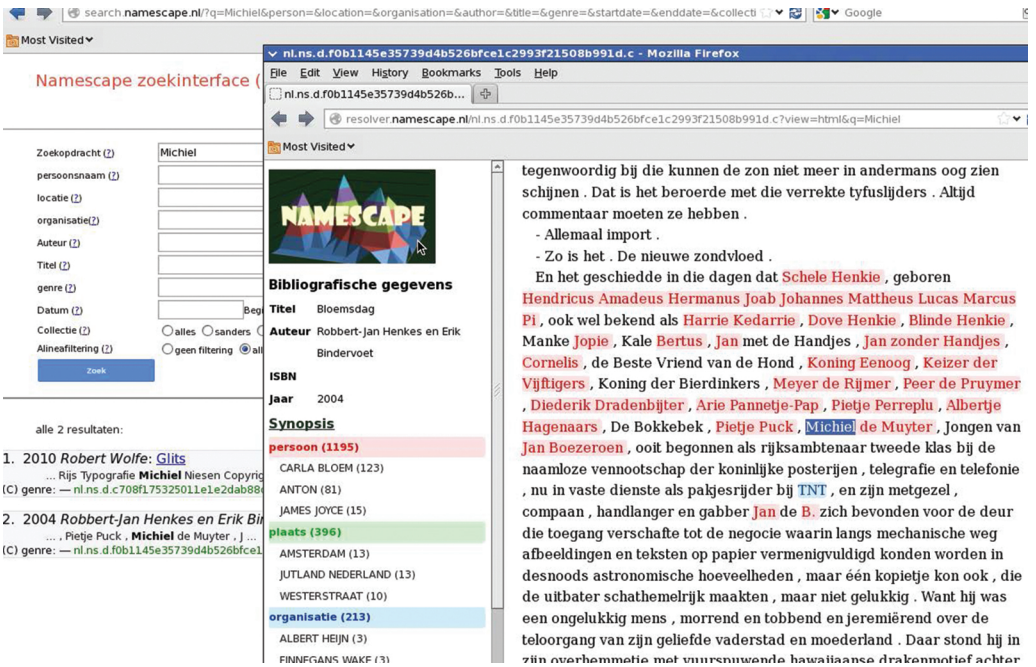

Figure 30.1: Components of the Namescape search interface. Left: search form (above) and hits (texts in which the name 'Michiel' occurs). Right: detailed view of the second hit, with an overview of all names (left) and part of the text with names highlighted (right). The three colours represent the three main name types: personal names (pink), place names (green, no example in the paragraphs shown), and names of organisations (blue). The example also shows that the tagger does not yield complete accuracy.

[8] Developed by Max Grim and Floris den Heijer under the supervision of Maarten Marx.

names (as recognised by an earlier tagger), with an automatically generated link to Wikipedia. The network of named entities in the novels is visualised in three ways: two different representations of the co-occurrence network, and a dispersion plot (see Figure 30.2).

Network of Characters Each book contains a network of named entities, usually characters. Two named entities are considered connected if they are both mentioned in the same paragraph. Clustering is performed according to the Louvain method (Blondel et al., 2008) for finding communities in social networks.

This is a fast algorithm that optimises a *modularity* criterion (the modularity of a partition is a measure that compares the density of links inside a cluster to links between clusters).[9]

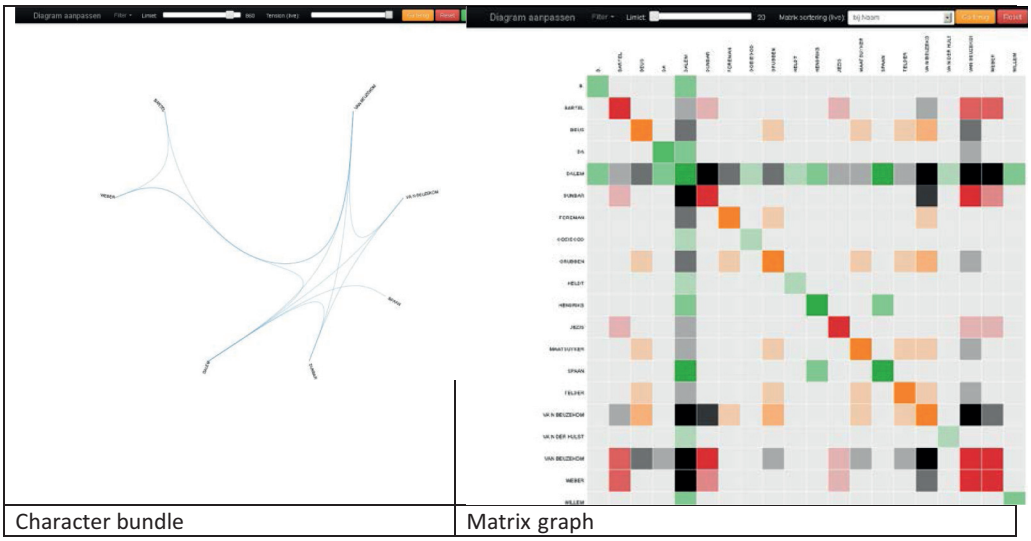

Figure 30.2: Examples of the visualisation options in Namescape for A. Alberts' *De vergaderzaal*. The Character bundle and Matrix graph are different visualisations of co-occurrence of names in the same paragraphs; the more co-occurrences, the thicker the lines (Character bundle) or the darker the colour (Matrix graph). The Dispersion graph shows in which paragraphs the names occur, linearly through the text from left to right.

[9] The algorithm proceeds by iterative application of two phases: a *modularity optimisation phase* which improves modularity by movement of nodes between clusters, and a *community aggregation phase* which builds a new network consisting of clusters resulting from the optimisation phase.

The character bundle and the matrix graph are different ways of displaying the network. The colours in the matrix graph correspond to the clusters, and the intensity of the colour indicates the frequency of the name co-occurrence in the book.

Dispersion Graph (Barcode Graph) The dispersion graph shows which character is mentioned in which paragraph. The horizontal axis represents paragraphs in the book, from the first on the left to the last on the right; the vertical axis represents characters. A coloured bar at (x, y) means that character y is mentioned in paragraph x. The dispersion measure (cf. Juilland et al., 1970), based on the frequency and the distribution of occurrences, is believed to be a good indicator for the prominence of a character in a novel (cf. Karsdorp et al., 2012 for this point in the context of folk tales).

Barcode Browser The Barcode Browser (http://barcode-browser.namescape.nl/index.xql) gives an overview of the search results for a collection of documents (see Figure 30.3). Each document in the search result is a column; the lines represent the paragraphs in the document. Paragraphs matching the search query are highlighted with a colour ranging from yellow (low relevance) to red (highly relevant).

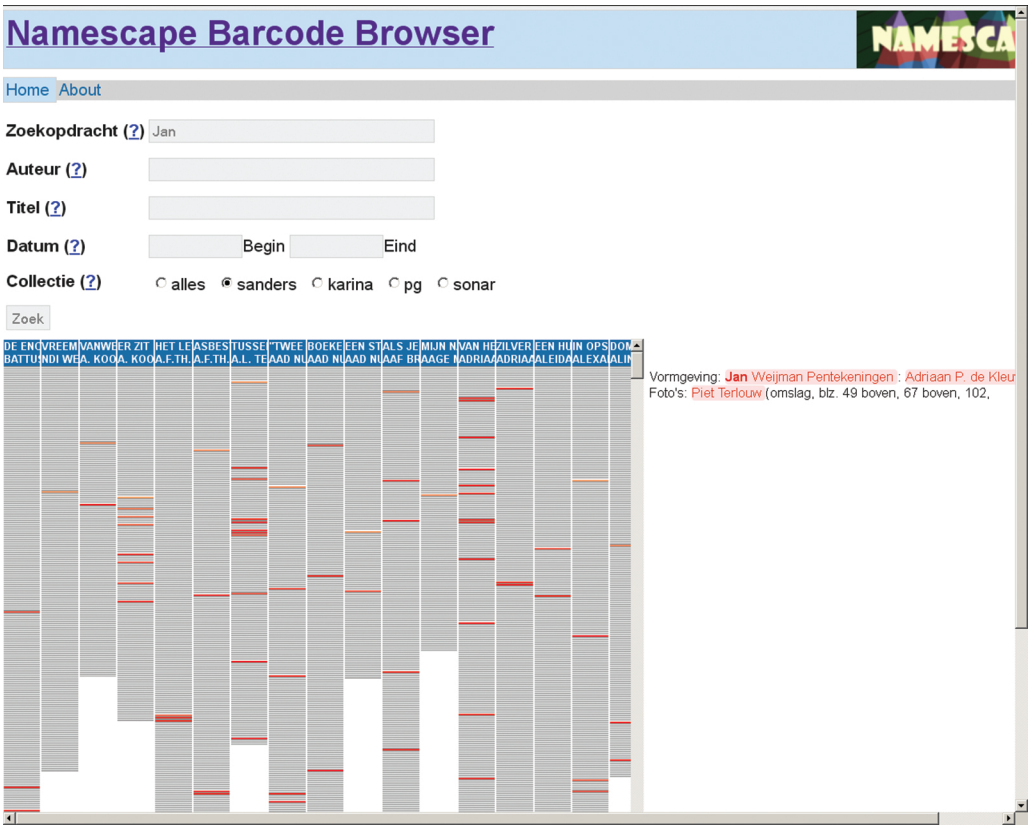

Figure 30.3: The Namescape Barcode Browser (part of the search interface). The example shows a search for the name 'Jan' in one of the subcorpora, with the search form above, and part of the results below. Each text in which the name was found is shown as a vertical column, with a bar for each paragraph. The paragraphs containing the name Jan are highlighted, from yellow (low relevance) to red (high relevance). Hovering over a coloured bar shows the text of the paragraph with all names in red and the search query in red and bold.

30.3 Evaluation

The most important results of the Namescape project are the specialised tagger for Dutch literary texts, and the visualisation tools to explore the landscape of names in individual texts. The search option is a great help to check the occurrence of certain names in a large corpus. Of course the corpus is still too small for really ambitious overviews, and it is unfortunate that we cannot make the full text of many works available, but it still gives a much broader view than ever before. Furthermore, although the scholar can get a nice overview on screen, it is not possible yet to download statistical reports or to view them in more detail. Another wish for the future is the option to submit privately owned texts to the tagger and download the results.

The Visualizer is a great tool to help explore the data. This is truly inspirational and may lead to many new hypotheses that could be tested in future projects. A drawback of the current Visualizer is that it does not have an upload function and only works on a static set of annotated novels which, furthermore, have been annotated with an older version of the NER tagger. There is therefore a need to update the tagging in the files underlying the Visualizer to make the explorations more reliable as well as more detailed (recall that the final Namescape NER distinguishes first names, family names and nicknames).

New onomastic research was not part of the Namescape project, but is in preparation. The new possibilities were not tested outside the project team. In a paper for the International Congress of Onomastic Sciences in Glasgow in August 2014, we gave an overview of the project, focusing on a problem many mainstream literary onomasticians still have when looking at the results of software. For the human eye, the results still show a lot of mistakes. Even now, many scholars conclude that it is therefore better not to use the tools at all. We think it would be better to find ways to deal with all this noise, and we described a couple of these potential solutions in an earlier paper (van Dalen-Oskam and de Does, 2016). Still, one of the ways we do not mention and which would certainly be useful is to try to improve the new tagger of literary texts.

30.4 What Came After

After the Namescape project, we have been able to enhance the performance of the SVM NER tagger by adding distributional word vectors (cf. Turian et al., 2010), produced by the word2vec program (Mikolov et al., 2013), as features to the classifier. This has yielded a significant improvement of tagging accuracy on the Namescape training corpus (cf. Table 30.2).

tagger	**NE type**	**precision**	**recall**	**F1**
Namescape SVM	location	0.83	0.729	0.776
	misc	0	0	0
	organisation	0.516	0.251	0.339
	person	0.867	0.917	0.896
	overall	0.853	0.838	**0.845**
Namescape SVM (+wv)	location	0.858	0.830	0.844
	misc	0.4	0.154	0.222
	organisation	0.656	0.459	0.54
	person	0.932	0.941	0.936
	overall	0.904	0.881	**0.893**

Table 30.2: Results of the enhancement of the Namescape tagger with distributional word embedding features.

The project also inspired a project called Beyond the Book; the ultimate aim of the researchers in this project was to examine if knowledge of names in a novel could contribute to a book being found interesting for readers in another language. They assumed that a novel that mentions a lot of culture-specific information may be less interesting for readers from other cultures, unless there is a special hype of literature focusing on the exotic. They thought that a tool that can show how exotic a novel is could be useful for publishers in helping them decide which novels to push for translation in which languages. To make a very first step towards this possible goal, Beyond the Book focused on names and applied the Semanticizer for named entity resolution. Names were linked to the most probable Wikipedia entry. For each of these Wikipedia entries the researchers calculated the number of contributors and their background (country of origin) and the number of edits. Then they compared these with the mean number of contributions from a certain country to the whole of Wikipedia. The difference between the outcome per entry then showed if the editors from certain countries made more changes than average to this entry, or less. If they made more, the scholars assumed that in the country of origin of these editors, the named entity was well-known and found culturally relevant. They explored several ways of visualising the results of such analysis for individual names and individual novels (Martinez-Ortiz et al., 2015). A tool that could be used by publishers and translators to get suggestions for the selection of novels for translation is still far away, however.

30.5 Future Work

Apart from finding more ways of dealing with noise, we have several other wishes for next steps in this research. Obviously, one would want to optimise the tools for automatic tagging. As we have seen, progress in the field of NE recognition is possible; for NE resolution, a first step is to develop more gold standard data. Furthermore, to truly turn the Namescape interactive environment into a virtual research environment that enables researchers to tag, explore, refine, and publish their data, we need to implement additional functionality.

After uploading documents to the NE tagger, researchers should be able to use the exploration and visualisation tools on their own data. To be able to deal with the noise problem described above, scholars should have the option to correct the markup after automatic tagging in a user-friendly way. Finally, we would like to have options to publish tagged material: users should at least be able to download not only their tagged texts (and, if there are no IPR issues at stake, to make them available to other users), but also statistical overviews of names and name co-occurrences.

References

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre 2008 Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), pp. P10008. http://stacks.iop.org/1742-5468/2008/i=10/a=P10008.

Chinchor, Nancy, Erica Brown, Lisa Ferro, and Patty Robinson 1999 1999 named entity recognition task definition, *Technical report* MITRE. ftp://ftp3.nist.gov/ace/phase1/ne99_taskdef_v1_4.ps.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning 2005 Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 363–370. http://dx.doi.org/10.3115/1219840.1219885.

Juilland, Alphonse, Dorothy Brodin, and Catherine Davidovitch [and Others] 1970 *Frequency Dictionary of French Words* Mouton.

Karsdorp, Folgert, Peter Van Kranenburg, Theo Meder, and Antal Van den Bosch 2012 Casting a spell: Identification and ranking of actors in folktales, *The Second Workshop on Annotation of Corpora for Research in the Humanities*, Lisbon, Portugal.

Landsbergen, Frank 2012 Evaluation of named entity work in IMPACT: NE Recognition and matching, *Technical report*.

Martinez-Ortiz, Carlos, Marijn Koolen, Floor Buschenhenke, Karina van Dalen-Oskam 2015 Beyond the Book: Linking Books to Wikipedia. 2015 IEEE 11th International Conference on eScience, Munich, Germany, p. 12–21.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean 2013 Efficient Estimation of Word

Representations in Vector Space, *Proceedings of Workshop at ICLR* Vol. abs/1301.3781.

Odijk, Daan, Edgar Meij, and Maarten de Rijke 2013 Feeding the second screen: Semantic linking based on subtitles, *Open research Areas in Information Retrieval (OAIR 2013)* Lisbon, Portugal.

Oostdijk, Nelleke, Martin Reynaert, Paola Monachesi, Gertjan Van Noord, Roeland Ordelman, Ineke Schuurman, and Vincent Vandeghinste 2008 From d-coi to sonar: a reference corpus for dutch., *LREC*.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman 2013 The Construction of a 500-Million-Word Reference Corpus of Contemporary Written Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch* Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg, pp. 219–247. http://dx.doi.org/10.1007/978-3-642-30910-6.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio 2010 Word representations: A simple and general method for semi-supervised learning, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 384–394. http://dl.acm.org/citation.cfm?id=1858681.1858721.

van Dalen-Oskam, Karina 2005 Vergleichende literarische Onomastik, *in* Brendler, A. and S. Brendler, editors, *Namenforschung morgen: Ideen, Perspektiven, Visionen* Baar, Hamburg, pp. 183–191. (An English translation 'Comparative Literary Onomastics', is available at https://www.huygens.knaw.nl/wp-content/bestanden/pdf_vandalenoskam_2005_Comparative_Literary_Onomastics.pdf)

van Dalen-Oskam, Karina 2013 Names in novels: an experiment in computational stylistics, *LLC: The journal of digital scholarship in the Humanities* 28 pp. 359–370. http://dx.doi.org/10.1093/llc/fqs007.

van Dalen-Oskam, Karina, Jesse de Does, Maarten Marx, Isaac Sijaranamual, Katrien Depuydt, Boukje Verheij, Valentijn Geirnaert 2014 Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal* 4 (2014), 121–136. 20 December 2014, http://www.clinjournal.org/sites/clinjournal.org/files/09-VanDalenOskam-etal-CLIN2014.pdf.

van Dalen-Oskam, Karina 2016 Corpus-based approaches to names in literature. Carole Hough (Ed.), *The Oxford Handbook of Names and Naming*. Oxford University Press, 2016, 344–354.

van Dalen-Oskam, Karina and Jesse de Does 2016 Namescape, or how to deal with noise. *Names and Their Environment' Proceedings of the 25th International Congress of Onomastic Sciences, Glasgow, 25–29 August 2014*. Eds. Carole Hough & Daria Izdebska, 5 Vols. Vol. 5, 57–64, http://www.icos2014.com/congress-proceedings/.