Computational Linguistics in the Netherlands Journal 4 (2014) 121-136 Submitted 06/2014; Published 12/2014

Named Entity Recognition and Resolution for Literary Studies

Karina van Dalen-Oskam*[*] KARINA.VAN.DALEN@HUYGENS.KNAW.NL
Jesse de Does[†] JESSE.DEDOES@INL.NL
Maarten Marx[‡] MAARTENMARX@UVA.NL
Isaac Sijaranamual[‡] ISAACSIJARANAMUAL@UVA.NL
Katrien Depuydt[†] KATRIEN.DEPUYDT@INL.NL
Boukje Verheij[†] BOUKJE.VERHEIJ@INL.NL
Valentijn Geirnaert[†] VALENTIJN.GEIRNAERT@INL.NL

[*] *Huygens Institute for the History of the Netherlands, Prins Willem-Alexanderhof 5, 2595 BE The Hague, The Netherlands*

[†] *INL, Matthias de Vrieshof 3, 2311 BZ Leiden, The Netherlands*

[‡] *ILPS Informatics Institute, Universiteit van Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.*

Abstract

This paper reports on the project Namescape: Mapping the Landscape of Names in Modern Dutch Literature, funded by CLARIN-NL. The background of the project is research in literary onomastics, the study of the usage and functions of proper names in literary (i.e. fictional) texts. The two main tasks for the project were to adapt existing Named Entity Recognition software to modern Dutch fiction, and to perform Named Entity Resolution by linking the names to Wikipedia entries. For Named Entity Recognition, existing tools have been trained on literary texts and a new NE tagger has been developed. The standard list of name categories had to be extended, since the analysis of the usage of proper names in literature needs to distinguish e.g. between first names and family names. The Named Entity Resolution task was done to explore the possibility of labeling the names in fiction in another way, by categorizing a name as referring to a person or location that only exist in the story of a fictional work (plot-internal names), or one referring to a person or location in the real world (plot-external names). This distinction is linked to the hypothesis that plot-internal and plot-external names can have different (stylistic and narrative) functions. Automatically marking them up is the first step toward testing that hypothesis on a large corpus. In this paper we describe the results of these two main tasks.

1. Introduction

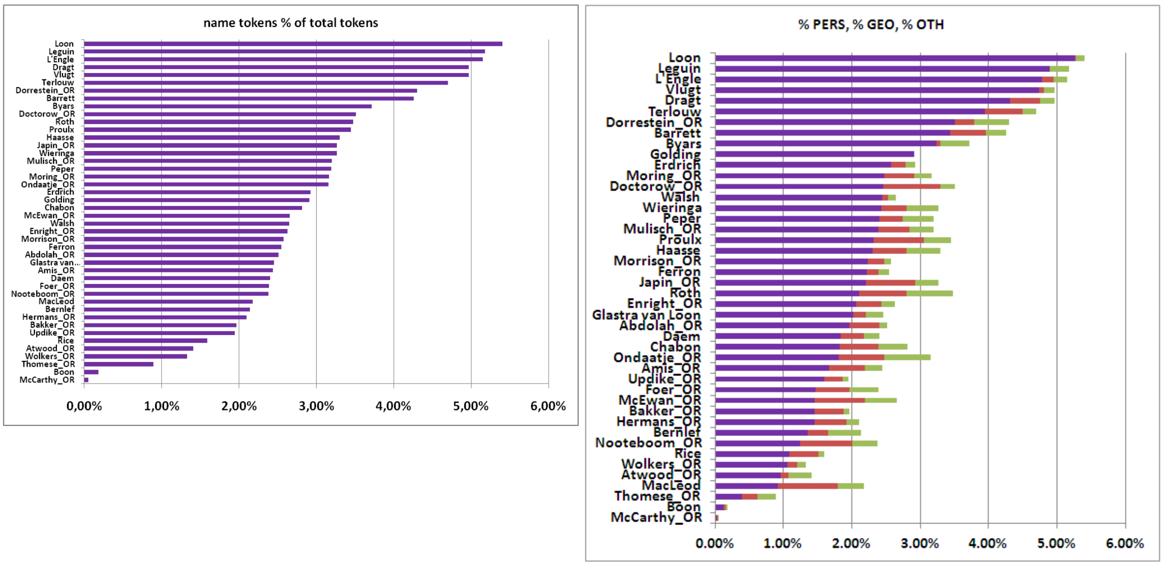
Research into proper names is not only carried out by linguists, but also by literary scholars. Because the questions asked by literary scholars partly differ from those asked by linguists, literary scholars also have different requirements for the tools they need for their work. This paper describes how the CLARIN-NL Demonstrator project [1] Namescape [2] developed data and tools suitable for comparative literary onomastics (name studies).

Comparative literary onomastics (van Dalen-Oskam 2005) has as its main research question: What is the usage of names in literary text and what are their (stylistic and narrative) functions? A simple analysis of the relative proportion of names versus non-names in a text and the proportion

[1] Demonstrator projects are projects that demonstrate the potential of software, in particular Human Language Technology, for the humanities. They do so by making a concrete application accompanied by a demonstration scenario to show how the application can be used for humanities research

[2] The project ran from May 2012 until the end of 2013 and was funded by CLARIN-NL. CLARIN-NL project code: 11.029.

Figure 1: Proportions (percentages) of name tokens and name types



(a) token level

(b) type level

of use of different types of named entities in typical literary works is a first step. For instance, the left pane of Figure 1 shows a bar chart with the percentage of tokens belonging to a proper name in each of the 44 novels in the pilot corpus, ranked from highest to lowest percentage. Most novels seem to have a percentage of around 2 to 3.5 percent. In the stacked bar chart on the right the same percentages are shown, but split into personal, geographical, and other name types. One can easily see that personal names are much more prominent than the other two name types. Other names occur in extremely low amounts, and geographical names are located somewhere in between. Starting from simple overviews like this, we can proceed to develop other quantitative indicators which can be used to draw conclusions about functions of names that could not be highlighted before in a verifiable and repeatable way.

Prior to the Namescape project, the first author performed a pilot analysis of 22 Dutch and 22 English novels (van Dalen-Oskam 2013), from which some hypotheses and observations emerged which prompted the need for a much larger research corpus. The ratio between the use of first names and family names may be indicative of the level of intimacy in a novel (van Dalen-Oskam 2005). Different functions can be assumed for plot-internal versus plot-external names; plot-external names refer to persons, places or objects outside the work of fiction (Obama, Buenos Aires, Lord of the Rings) and they seem mostly used as characterizations of the plot-internal (fictional) characters. Different novels and authors etc. make a different use of these plot-external names and for different reasons. Furthermore, quantitative study of the pilot corpus led to the identification of two functions of geographical names that have not been noticed before: the high use of a large number of distinct geographical names embodied a geographical taboo in a novel, and the articulation of these lists of names functioned as a calming mantra for the main character.

One of the conclusions from the exploration of the pilot corpus is that it would, for example, be extremely useful to be able to get a quick overview of the ratio of plot-internal and plot-external names in a large corpus of novels, to learn more about their possible functions. To proceed along these lines, literary scholars require at least two things. First, a much larger corpus of literary works. Second, a set of tools to assist the researcher to analyze this larger corpus. Such were the goals of Namescape project: to acquire and curate a large corpus of modern dutch literature, to develop new tools for named entity tagging and resolution of literary works, allow search and provide insightful visualizations. The tagged corpora and the annotation tools enable literary scholars to perform research on a much larger annotated corpus than would have been possible to do manually; the exploratory functions of the visualization tools in the demonstrator are expected to lead to many more new observations, questions, and inspirations. However, the switch from manual data annotation to the use of automatic tools is not without problems. How reliable are our conclusions when the errors made by the tools cause a considerable amount of noise?

The outline of this paper is as follows. Section 2 will describe the components of the Namescape research environment (the corpora, named entity annotation, tagging, resolution, search interface and visualization tools.) In section 3 we will evaluate the suitability of the tools for the literary onomast, followed by future work in section 4.

2. Namescape research environment components

This section describes the components of the Namescape research environment.

2.1 Namescape Corpora

Namescape has developed two corpora with manual annotations: Corpus Huygens and Corpus Namescape. Corpus Huygens (˜1.5M tokens) consists of the 22 Dutch novels which were also used for the pilot project. It is the only corpus for which all NE annotations have been manually verified. Corpus Namescape (˜28M tokens) consists of a aset of 550 Dutch books from the period 1970–2009. These books have been scanned and OCRed from the originals, the OCR results have not been further corrected.

The manually annotated *Namescape gold standard corpus* consists of random paragraphs from the core Namescape corpus. The reason for choosing our training data in such a way is twofold: On the one hand, if the training corpus had contained complete novels, it would not be distributable due to IPR problems, whereas the current gold standard corpus can be considered a set of extended quotations. Moreover, choosing training data in this way has a positive effect on the machine learning procedures.

Three additional corpora have been collected and curated for Namescape: Corpus eBooks, Corpus Sonar and Corpus Gutenberg. Corpus eBooks consists of over 7000 books (˜500M tokens), all in EPUB[3] format. Although all of them are Dutch, a considerable number of these are books translated English works. It is also the most varied corpus in genre and register. Corpus Sonar (˜11M tokens) consists of just over 100 books from the SoNaR-corpus (Oostdijk et al. 2008). Corpus Gutenberg (˜30M tokens) are the 530 Dutch books from Project Gutenberg.[4] The books in this corpus are mainly from the 17th to 20th century, so it is the only corpus which contains historical text. The Gutenberg books have been manually transcribed, and contain material in historical spelling, which makes NER significantly more difficult.

2.2 Annotation scheme

Text structure annotation

We have adopted the widely used TEI P5 standard[5] for the XML representation of corpus text structure and we have developed a simple extension to TEI to tag the named entity properties needed in a way that accommodated our need to be able to query the tagged data in a simple way. For easy querying, we preferred all searchable properties of names to be 'inline' rather than standoff.

[3] http://idpf.org/epub

[4] Project Gutenberg's mission statement is: to encourage the creation and distribution of eBooks (http://www.gutenberg.org/).

[5] The Text Encoding Initiative, a standard for the representation of texts. http://www.tei-c.org/Guidelines/P5

This entails the introduction of several additional attributes for tagged named entity occurrences. We also preferred to use one single tag for named entities, and one (different) tag for entity parts, whereas TEI offers the choice of using either a range of tags (`persName`, `geoName`, `orgName`), or the single tag `name`, which latter choice is inconvenient for querying because tagging name parts would lead to nested `name` tags:

```
<name type="person">
  <name type="forename">Jan</name>
  <name type="surname">Janssen</name>
</name>
```

We will not describe the annotation scheme in detail here, but a few examples should give an impression.[6] The examples shown here use the full annotation scheme which are the result of the conversion of the manual annotations. The annotations produced by the automatic taggers is slightly simpler in that it omits the attributes `resolution` and `gloss`. These were not considered feasible.

The `ne` element and the `nePart` element mark entity references in running text:

```
<ns:ne
   xmlns:ns="http://www.namescape.nl/"
   type="person"
   gloss="MAIN CHARACTER"
   structure="forename"
   nymRef="nym7"
   normalizedForm="MICHIEL"
   resolution="plotInternal">
       <ns:nePart type="forename" sex="male">Michiel</ns:nePart>
 </ns:ne>
```

The header contains a `listNym` element which summarizes the named entity use in the text:

```
<nym ns:id="nym7" ns:resolution="plotInternal" ns:gloss="MAIN CHARACTER" ns:type="person">
  <usg type="frequency">531</usg>
  <form type="nym">MICHIEL VAN BEUSEKOM</form>
  <form type="witnessed">
    <orth type="original">Michiel</orth>
    <orth type="normalized">MICHIEL</orth>
    <usg type="frequency">501</usg>
  </form>
  ...
  <form type="witnessed">
    <orth type="original">v.B.</orth>
    <orth type="normalized">V.B.</orth>
    <usg type="frequency">1</usg>
  </form>
</nym>
```

NE annotation guidelines

We have followed the 1999 Named Entity Recognition Task Definition (Chinchor et al. 1999) for the basic principles of named entity annotation. This means that the gold standard data developed in the project follows the sample principles as other important data sets developed for Dutch.

2.3 Named entity tagging for Dutch Novels

Dutch is not an under-resourced language with respect to named entity tagging. Training and evaluation corpora for NER are available, and multiple NE taggers have been trained by means of supervised machine learning.

[6] The entire annotation scheme is described in (Namescape 2013)

1. The CoNLL 2002 shared task (Tjong Kim Sang 2002) website provides test and training corpora.[7] The Dutch task can be considered difficult; best results do not exceed an F1 score of about 80%.
2. In the SONAR (500-million-word reference corpus of contemporary written Dutch) project (Oostdijk et al. 2013) a NER training corpus of about 1m tokens has been developed.[8]
3. In the IMPACT[9] project, gold standard data and an extension of the Stanford NE tagger have been developed for historical Dutch (Landsbergen 2012).[10]

Namescape NER tagging

We have deployed two distinct named entity taggers trained on the Namescape training corpus. The first one is the Conditional Random Field-based Stanford tagger,[11] with the default settings. The second is a Support Vector Machine-based tagger,[12] which has been designed to improve performance by making use of information derived in an unsupervised way from a corpus (in this case the core Namescape corpus).

Comparing these two tagger architectures,[13] it should be mentioned that with identical feature sets, the proper sequence model of the CRF-based Stanford tagger typically outperforms the SVM-based tagger, which performs classification per-token (it is not a sequence model.) This is in accordance with the observations in (Li et al. 2008) and (Desmet and Hoste 2014). The latter moved from an earlier SVM-based classifier to a CRF-based one.

However, using the SVM classifier permits the inclusion of continuous-valued feature sets, which turn out to enhance the performance to a level above that of the default Stanford setup. After the Namescape project, we have been able to enhance the performance 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.[14]

Evaluation

Evaluation of entity-level performance of both taggers has been performed using a fixed 90%/10% split of the gold standard into training data and evaluation data. The results are presented in Table 1.

[7] The data sets and evaluation software can be found at `http://www.cnts.ua.ac.be/conll2002/ner/` The Dutch data consist of four editions of the Belgian newspaper "De Morgen" from 2000 (June 2, July 1, August 1 and September 1). The data was annotated as a part of the *Atranos* project at the University of Antwerp. It consists of a training part of about 200k tokens and two test sets of 38k and 69k tokens respectively.

[8] The annotated corpus was used for the development of a NE classifier (Desmet and Hoste 2010), which was used for the automatic annotation of the remaining 499 million words. The best result obtained for main entity type classification is an F1 score of 84.44%, which was realized with an ensemble of classifiers. The best individual classifier lagged only 0.67% behind that (Desmet and Hoste 2014).

[9] `http://www.digitisation.eu`

[10] It is a mixed corpus of 18th, 19th and early 20th century data with about 900k tokens from the "Digitale Bibliotheek voor de Nederlandse Letteren" (`http://www.dbnl.org`), 254k tokens from the Dutch Royal Library historical newspaper collection (`http://kranten.kb.nl`) and about 500k tokens from the Dutch Parliamentary Papers collection (`http://www.statengeneraaldigitaal.nl`). NER results vary from around 80% on newspaper and Parliamentary papers to about 60% on the DBNL data.

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

[12] The Namescape 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/`.

[13] We cannot go into the technicalities of the two machine learning procedures for which we refer to the literature, cf. for instance (Sutton and McCallum 2012) and (Cristianini and Shawe-Taylor 2000). The relevant differences, from our point of view, are that 1) The CRF optimizes the probability of a label assignment over a sentence, whereas the SVM proceeds on a token-by-token basis; 2) The Stanford CRF can only incorporate discrete information, whereas the SVM may base its decisions on real-valued parameters.

[14] Similar improvements can also be observed on the CoNLL and Sonar corpora.

Table 1: Results of tagger evaluation on the Namescape training corpus. Stanford is the CRF based tagger, Namescape is our newly developed SVM based tagger and Namescape (+wv) adds the distributional word vectors.

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
Namescape (+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

These results look decent at first sight, but researchers are often disappointed by the error rate. We will deal with this issue in the section 3.2.

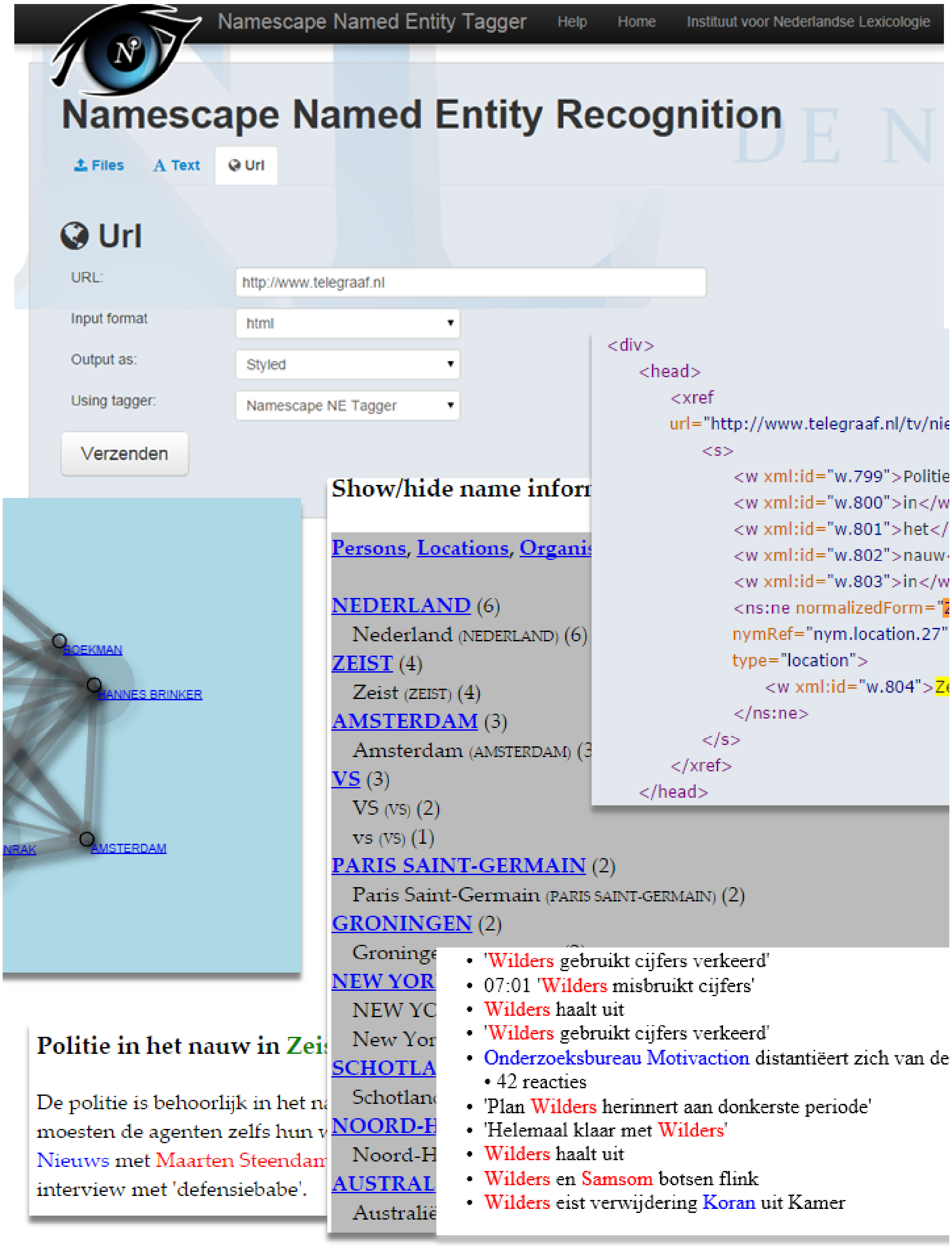
Web service and application

Namescape has delivered two tools for named entity recognition that users can apply to their own digital texts. A web service can be invoked by a program or script to add named entity tagging to a document, and a more user-friendly web interface shown in Figure 2. Both are available at `http://ner.namescape.nl/namescape/tagger` and allow submitting a text in one of several formats: plain text, HTML, EPUB, Word, TEI. Files can be uploaded 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. It can either be delivered as-is to the user (the 'raw' output option), or formatted and displayed with the names 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.

2.4 Named Entity Resolution

NE resolution has been performed using the ILPS semanticizer (Odijk et al. 2013), cf. `http://semanticize.uva.nl/doc/`. The semanticizer performs a task known as wikification, or entity linking, which means that it will find likely references to entities (represented by Wikipedia articles) in running text. This process is performed in two steps. First, entity mention candidates are identified. The candidate mentions are all contiguous n-grams which are also Wikipedia article titles or anchor texts to these articles. Second, for each candidate mention the candidate entities are ranked based on the sense probability, the score that given mention refers to that entity. The sense probability is determined by taking the product of link probability, the probability that the

Figure 2: Web application for NE tagging of Dutch Novels



Namescape Named Entity Tagger
Help
Home
Instituut voor Nederlandse Lexicologie
Files
Text
Url
Verzenden
BOEKMAN
HANNES BRINKER
NRAK
AMSTERDAM

Namescape Named Entity Recognition

Url

URL: http://www.telegraaf.nl

Input format: html

Output as: Styled

Using tagger: Namescape NE Tagger

```
<div>
   <head>
      <xref
      url="http://www.telegraaf.nl/tv/nie
         <s>
            <w xml:id="w.799">Politie
            <w xml:id="w.800">in</w
            <w xml:id="w.801">het</
            <w xml:id="w.802">nauw
            <w xml:id="w.803">in</w
            <ns:ne normalizedForm="Z
            nymRef="nym.location.27"
            type="location">
               <w xml:id="w.804">Ze
            </ns:ne>
         </s>
      </xref>
   </head>
```

Show/hide name infor

Persons, Locations, Organis

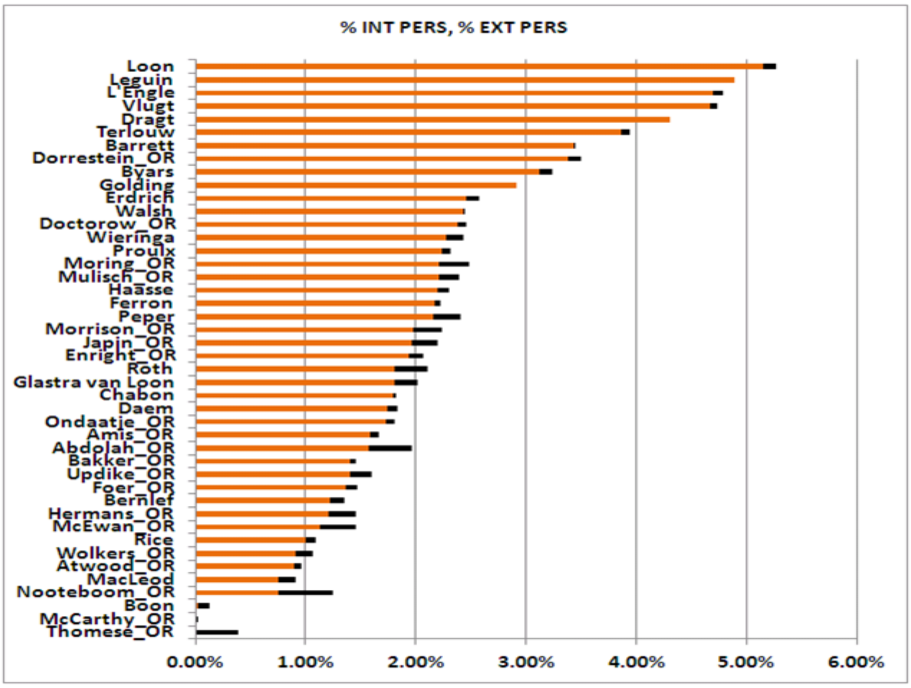
NEDERLAND (6)
Nederland (NEDERLAND) (6)
ZEIST (4)
Zeist (ZEIST) (4)
AMSTERDAM (3)
Amsterdam (AMSTERDAM) (3
VS (3)
VS (VS) (2)
vs (VS) (1)
PARIS SAINT-GERMAIN (2)
Paris Saint-Germain (PARIS SAINT-GERMAIN) (2)
GRONINGEN (2)
Groninge
NEW YOR
NEW YC
New Yor
SCHOTLA
Schotlanc
NOORD-H
Noord-H
AUSTRAL
Australië

Politie in het nauw in Zeis

De politie is behoorlijk in het na moesten de agenten zelfs hun v Nieuws met Maarten Steendam interview met 'defensiebabe'.

- 'Wilders gebruikt cijfers verkeerd'
- 07:01 'Wilders misbruikt cijfers'
- Wilders haalt uit
- 'Wilders gebruikt cijfers verkeerd'
- Onderzoeksbureau Motivaction distantiëert zich van de
  • 42 reacties
- 'Plan Wilders herinnert aan donkerste periode'
- 'Helemaal klaar met Wilders'
- Wilders haalt uit
- Wilders en Samsom botsen flink
- Wilders eist verwijdering Koran uit Kamer

Figure 3: Proportions of plot-internal versus external entity mentions



given mention is a link to an entity, and the prior probability, the probability the entity is linked to by this mention. This means that we do not take the context of the occurrences into account.

Since the named entity mentions have already been tagged in the texts, we only apply the second step to the mentions found by the NE-tagger. The plot-externalness was then determined as follows: a given named entity is categorized as plot-external if it is resolved to an article which is about a non-fictitious entity.[15]

In the manually tagged pilot corpus of 44 novels, plot-internal names are much more prominent than plot-external names, as can be seen in Figure 3.

Evaluation

A real evaluation of the resolution accuracy is currently impossible: we simply do not have the gold standard data required. Fortunately, in the Corpus Huygens the entities have been manually tagged with their plot-internalness and -externalness which we can use for an indirect evaluation. With 3862 distinct name types and 35852 name tokens, we obtain a type accuracy of 74.5% and a token accuracy of 79.8% for the automatic tagging of this distinction by marking an entity as plot-external if and only if it could be resolved to a Dutch Wikipedia article.

2.5 Search interface

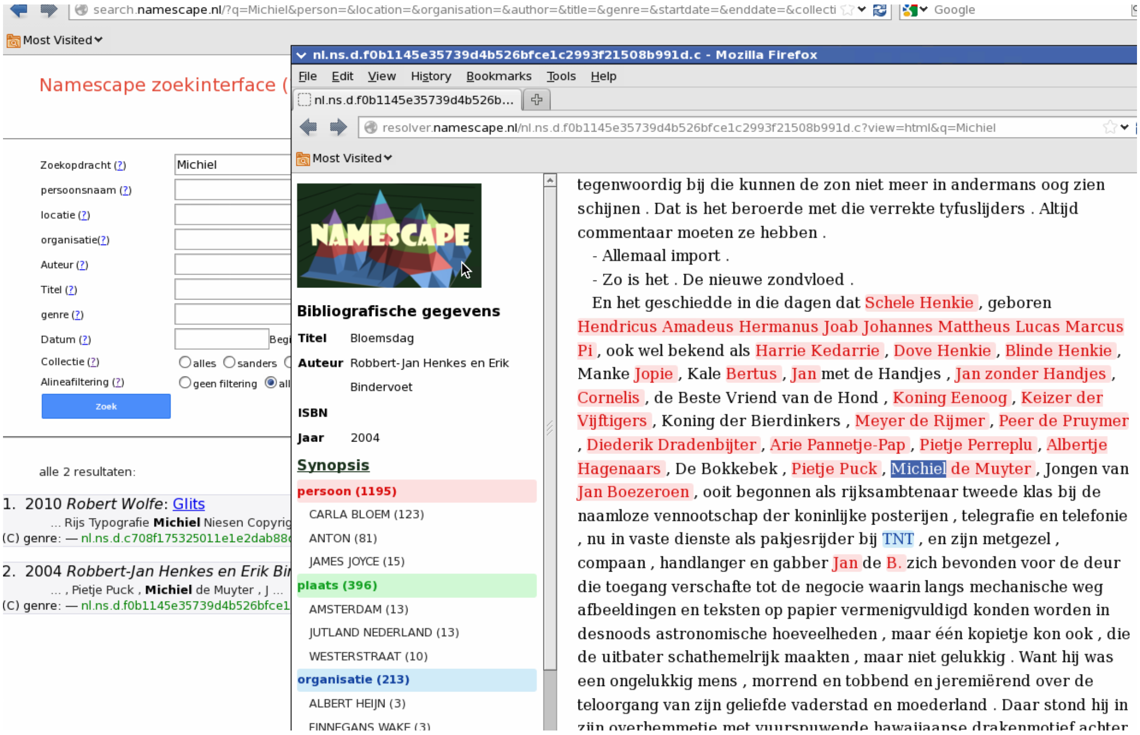
A search interface to the Namescape is available at `http://search.namescape.nl`. It is implemented in XQuery using the eXist XML database[16] and it enables full text search, searching and filtering by metadata and styled display of full text of the result. The publicly visible version at the abovementioned address is identical to the private installation, but access to the full texts is

[15] The fictitious features we used are: the article title or category contains any variant of ‘legend’, ‘mythological’ or ‘fictional’.

[16] `http://www.exist-db.org`

restricted to a subset of the corpora, exactly those works for which the IPR allow any access to the data at all: Corpus Gutenberg and Corpus Sonar.

Figure 4: Search interface on the left and highlighted full text result on the right.


search.namescape.nl/?q=Michiel&person=&location=&organisation=&author=&title=&genre=&startdate=&enddate=&collecti
Google
Most Visited
nl.ns.d.f0b1145e35739d4b526bfce1c2993f21508b991d.c - Mozilla Firefox
File Edit View History Bookmarks Tools Help
nl.ns.d.f0b1145e35739d4b526b...
resolver.namescape.nl/nl.ns.d.f0b1145e35739d4b526bfce1c2993f21508b991d.c?view=html&q=Michiel
Most Visited
NAMESCAPE

Namescape zoekinterface (

Zoekopdracht (?) Michiel
persoonsnaam (?)
locatie (?)
organisatie(?)
Auteur (?)
Titel (?)
genre (?)
Datum (?) Begi
Collectie (?) ○ alles ○ sanders
Alineafiltering (?) ○ geen filtering ● all

Zoek

alle 2 resultaten:

1. 2010 *Robert Wolfe*: Glits
… Rijs Typografie **Michiel** Niesen Copyrig
(C) genre: — nl.ns.d.c708f175325011e1e2dab88

2. 2004 *Robbert-Jan Henkes en Erik Bi*
… , Pietje Puck , **Michiel** de Muyter , J …
(C) genre: — nl.ns.d.f0b1145e35739d4b526bfce1

Bibliografische gegevens

Titel Bloemsdag
Auteur Robbert-Jan Henkes en Erik Bindervoet
ISBN
Jaar 2004

Synopsis

persoon (1195)
CARLA BLOEM (123)
ANTON (81)
JAMES JOYCE (15)
plaats (396)
AMSTERDAM (13)
JUTLAND NEDERLAND (13)
WESTERSTRAAT (10)
organisatie (213)
ALBERT HEIJN (3)
FINNEGANS WAKE (3)

tegenwoordig bij die kunnen de zon niet meer in andermans oog zien schijnen . Dat is het beroerde met die verrekte tyfuslijders . Altijd commentaar moeten ze hebben .

- Allemaal import .

- Zo is het . De nieuwe zondvloed .

En het geschiedde in die dagen dat Schele Henkie , geboren Hendricus Amadeus Hermanus Joab Johannes Mattheus Lucas Marcus Pi , ook wel bekend als Harrie Kedarrie , Dove Henkie , Blinde Henkie , Manke Jopie , Kale Bertus , Jan met de Handjes , Jan zonder Handjes , Cornelis , de Beste Vriend van de Hond , Koning Eenoog , Keizer der Vijftigers , Koning der Bierdinkers , Meyer de Rijmer , Peer de Pruymer , Diederik Dradenbijter , Arie Pannetje-Pap , Pietje Perreplu , Albertje Hagenaars , De Bokkebek , Pietje Puck , Michiel de Muyter , Jongen van Jan Boezeroen , ooit begonnen als rijksambtenaar tweede klas bij de naamloze vennootschap der koninlijke posterijen , telegrafie en telefonie , nu in vaste dienste als pakjesrijder bij TNT , en zijn metgezel , compaan , handlanger en gabber Jan de B. zich bevonden voor de deur die toegang verschafte tot de negocie waarin langs mechanische weg afbeeldingen en teksten op papier vermenigvuldigd konden worden in desnoods astronomische hoeveelheden , maar één kopietje kon ook , die de uitbater schathemelrijk maakten , maar niet gelukkig . Want hij was een ongelukkig mens , morrend en tobbend en jeremiërend over de teloorgang van zijn geliefde vaderstad en moederland . Daar stond hij in zijn overhemmetje met vuurspuwende hawaiiaanse drakenmotief achter

Barcode browser

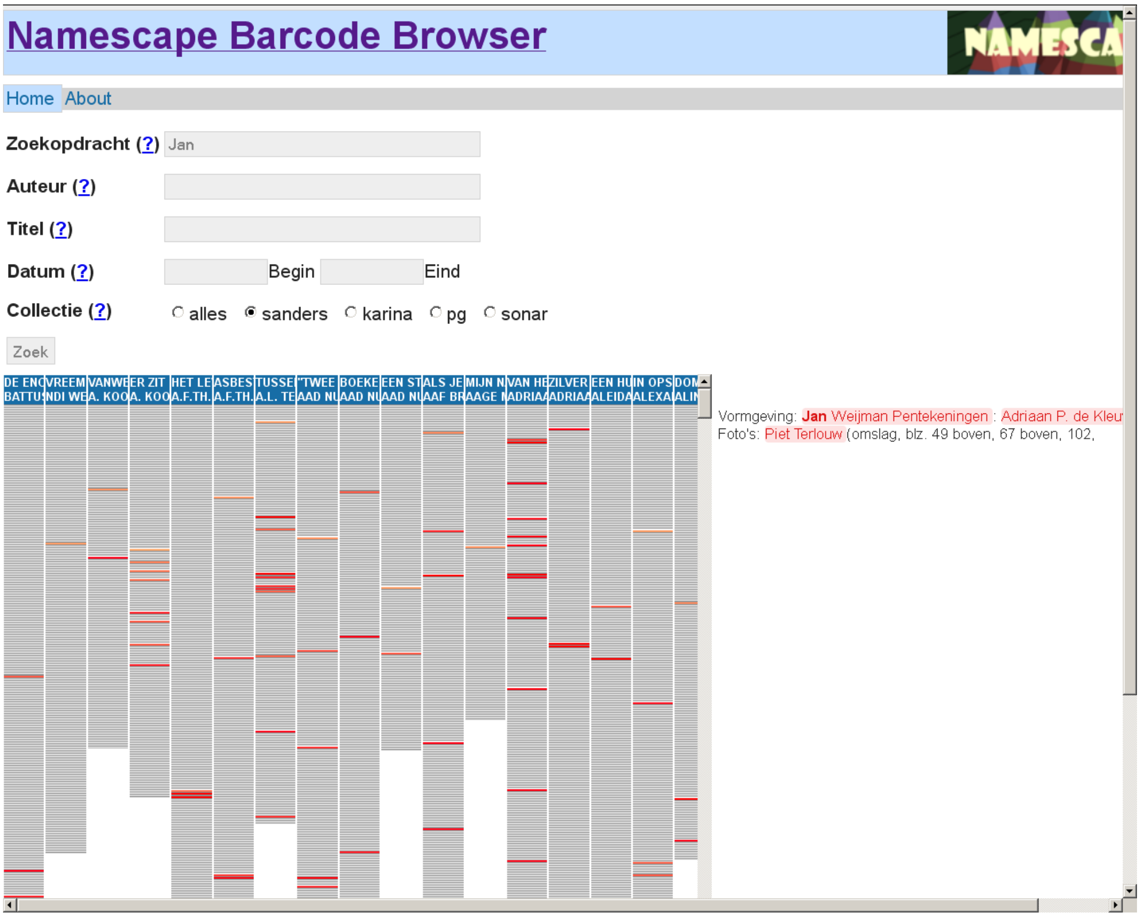
The Barcode Browser provides an alternative visualization of search results. It is a bird's-eye view of the search results for a collection of documents. Each document satisfying the filter (if any) is now represented by a column. The lines in the columns represent the paragraphs in that document. Paragraphs that match the search query are highlighted. The color used to highlight the paragraph ranges from yellow, for low relevance, to red, for highly relevant paragraphs. This gives a quick overview not only which documents match, but also where in the documents they match and this can easily be compared between documents.

The barcode browser for the Namescape corpus can be accessed at `http://barcode-browser.namescape.nl`.

2.6 Visualizations

The Namescape Visualizer was developed by Max Grim and Floris den Heijer under the supervision of Maarten Marx. The Visualizer helps to give an overview of the names found in a text and shows the co-occurrence of names in paragraphs. Suppose one would like to know the onymic landscape, the "landscape" of proper names, in the novel "De vergaderzaal" by A. Alberts. One then goes to `http://visualizer.namescape.nl/`, chooses "Bekijk alle boeken" ("browse all books"), and subsequently "Auteur" ("author"). Clicking on the A. Alberts novel "De vergaderzaal" results in an overview of the top twenty most frequent names (as recognized by an earlier tagger), with an automatically generated link to Wikipedia (`http://visualizer.namescape.nl/book/508`). The

Figure 5: Barcode browser



Namescape Barcode Browser

Zoekopdracht (?) Jan

Auteur (?)

Titel (?)

Datum (?) Begin Eind

Collectie (?) ○ alles ● sanders ○ karina ○ pg ○ sonar

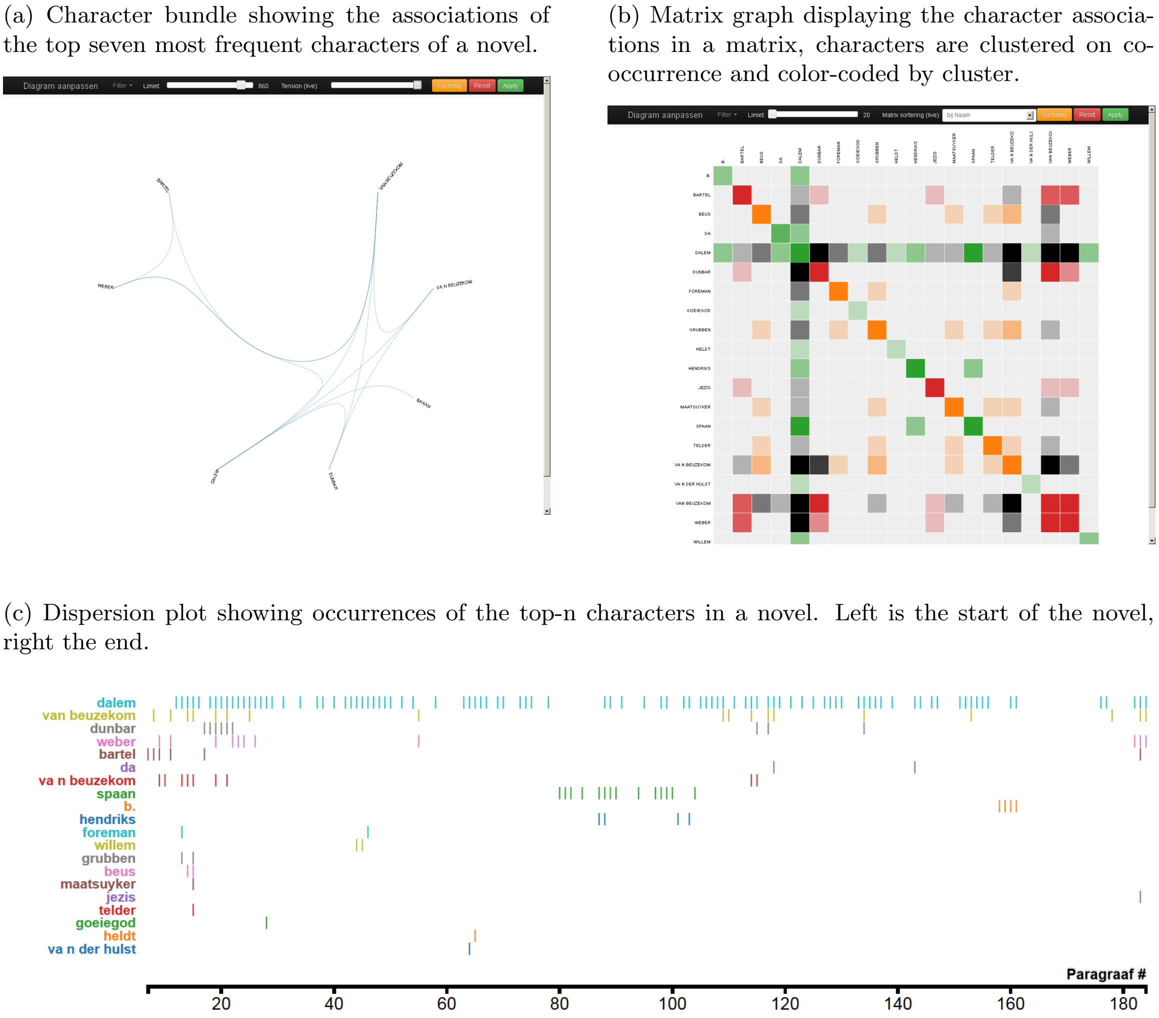
Vormgeving: Jan Weijman Pentekeningen : Adriaan P. de Kleu
Foto's: Piet Terlouw (omslag, blz. 49 boven, 67 boven, 102,

network of named entities (mostly characters, which usually occur with the highest frequencies) in the novels is visualized in three ways: two different representations of the co-occurrence network, and a dispersion plot, which can all be accessed at the bottom of this overview page.

Network of characters

Each book contains a network of named entities (mostly indicating characters). Two named entities are considered connected if they are both mentioned in the same paragraph. In the resulting graph, strongly interconnected clusters of the network can be identified. The character bundle and the matrix graph are different ways of displaying the network. The user can order the characters by name, frequency or by cluster. The clustering has been performed according to the Louvain method (Blondel et al. 2008). The colors in the matrix graph correspond to the clusters, the intensity of the color indicates frequency of the name in the book. See for example the two visualizations of the novel De vergaderzaal by A. Alberts: the bundle `http://visualizer.namescape.nl/graph/index/type/bundle/bookid/508` and the matrix graph `http://visualizer.namescape.nl/graph/index/type/matrix/bookid/508`.

Figure 6: Namescape visualizations



(a) Character bundle showing the associations of the top seven most frequent characters of a novel.

(b) Matrix graph displaying the character associations in a matrix, characters are clustered on co-occurrence and color-coded by cluster.

(c) Dispersion plot showing occurrences of the top-n characters in a novel. Left is the start of the novel, right the end.

Dispersion graph (barcode graph)

The dispersion graph shows which character is mentioned in which paragraph. The characters in the plot are sorted (in descending order) by dispersion. The horizontal axis represents paragraphs in book order, the vertical axis corresponds to characters. A colored bar at (x, y) means that character y is mentioned in paragraph x. The dispersion measure[17], based on the frequency and the distribution of occurrences, seems a good indicator for the prominence of a character in a novel (cf. (Karsdorp et al. 2012) in the context of folk tales).

Dispersion plots are not unlike the more artistic narrative charts.[18] The example from A. Alberts's De vergaderzaal `http://visualizer.namescape.nl/graph/index/type/barcode/bookid/508` is illuminating in this respect.

3. Evaluation from the researcher's point of view

How useful are the current tools? It is clear that NER performance is still in many cases unsatisfactory. Resolution, in turn, suffers from the inaccuracies of the NER system, and still leaves much to be desired. On the other hand, the scholar is able to explore much more data than ever before. Based on these explorations, a researcher can select those novels that need special attention, and zoom in on them using all the tools the Namescape demonstrator has delivered.

3.1 Prominent error types

This subsection describes the most frequent types of error introduced by the automatic tagging tools.

Named Entity Recognition

Tagging of the MISC and ORG entity classes is very poor, which is probably due in part to the low frequency of occurrence of these categories in the training data. Despite the corpus-based word representations which improve results on names not occurring in the training corpus, tagging of completely unknown (i.e. they do not occur in the background corpus from which the word representations have been obtained) entities is still rather poor. The word representations should probably be adaptive, in such a way that material to be processed can be added dynamically. Especially annoying are inconsistencies within one novel: the same word may get differently tagged without any obvious reason.

Named Entity Resolution

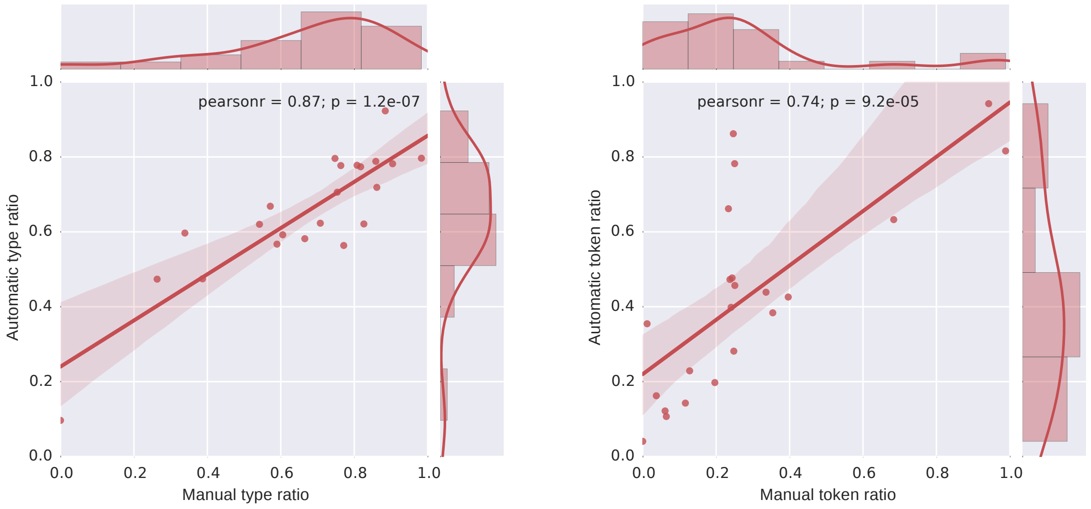
Often, a link cannot be presented, and many times a given link does not lead to the correct entry. The most prominent type of error is perhaps *over-resolution*: isolated first name or surname parts referring to a plot-internal character are often resolved to an apparently unconnected Wikipedia entry. This is understandable when we take into account that most proper names in a novel are expected to be plot-internal. Moreover, one has to take into account that the semanticizer has been designed to optimize the choice between different possible resolutions, rather than the decision between resolution and non-resolution, which is basically what we are interested in here.

However, the current links are useful, even if they do not lead to the correct entry. Even a wrong link can signal that a name is somehow recognizable, in a certain cultural context as described in the Wikipedia entry. When the name "van der Horst" from the novel "De vergaderzaal" by A. Alberts wrongly leads to the Wikipedia entry about a historical person named Ewoud Pietersz van der Horst,

[17] cf. (Juilland et al. 1970), which defines the dispersion of a word in a text consisting of n chunks (here: paragraphs) as $D = 1 - \frac{V}{\sqrt{n-1}}$, where the variation coefficient V is defined by $V = \frac{\sigma}{\mu}$, σ is the standard deviation of the frequencies per chunk, and μ is the mean frequency.

[18] cf. `http://xkcd.com/657/`

Figure 7: Ratio of proportions of tagged plot-external entities (manual vs. automatically tagged.)



(a) type level

(b) token level

`http://nl.wikipedia.org/wiki/Ewoud_Pietersz_van_der_Horst`, the entry may not refer to the fictional character of the Dutch novel, but it does correctly show that the name is from the Dutch language area. This approach will be further explored in a new project led by the first author, called "Beyond the Book", a one-year project funded by the eScience Center (NLeSC) which started May 1st 2014. The main research question there is if we can predict international accessibility of a novel based on textual features such as proper names.

One lesson we have learned from the project is that *internal entity resolution* or *entity co-reference* within a novel is an important first issue to be tackled before external resolution. The current NE tagger implements a rather crude heuristic to group different mentions of the same character. A better strategy is expected to help avoid many erroneous external resolutions.

3.2 Dealing with noise

In the overviews from the pilot corpus, the first author of this paper studied indicators like the percentage of tokens that are (part of) of a proper name, and the percentage of plot-internal versus plot-external entities as characteristics of a work. What happens when we attempt to compute such indicators from the output of the tools?

As an example, we try to obtain the internal/external ratio automatically from the output of the NE resolution process, by hypothesizing that an entity is plot-external if and only if the resolver comes up with a Wikipedia link. As we have seen, this is by no means warranted, cf. the preliminary evaluation on name type and token level in subsection 2.4, but for the moment, it is our best guess. On the pilot corpus, we can then compare this to the rate obtained from the manual tagging of the internal/external distinction.

In Figure 7a the 22 novels from the pilot corpus are arranged by *true* internal/external entity ratio, computed on the type level. The solid curve refers to this manually tagged quantity; the dashed curve gives the rate computed automatically. As one can see, the result is not perfect, but the correlation between the manual and automatic ratios is clear (Pearson $r = 0.87$).

This means that we can try to use this statistic to explore new corpora in a meaningful way. For instance, the titles with the lowest degree of external resolution in the eBooks corpus mostly have a distinct fantasy ring to them[19]:

- Julia
- Wolfsblad
- Meester Magier - vijfde zwaard - Rastoth
- Rhialto De Schitterende
- Wolfsblad Voor Altijd
- De Vloedvormer
- Meester Magier - vierde zwaard - Fiander
- In dienst van de Godin
- Damin Wolfsblad
- De Scrypturist
- De Bruiden Van Tyobar
- Meester Magier - eerste zwaard - Sperling
- Winter In Eden
- Cugel Gewroken
- vuurcyclus 3 - Schaduwvaan
- De dwergen deel 1
- De Zwarte Trillium
- De getijden van Blenholme

It seems, however, that on the token level the automatic tagger may show much more erratic behavior: the correlation is less pronounced ($r = 0.74$), but a pattern can still be detected: externalness on the token level is systematically overestimated (Figure 7b).

A systematic comparison, for a representative sample of the manually tagged novels compared and the automatically tagged versions of the same text, could help further analysis. One could examine the relation between the estimated values and the true values more closely, e.g. by calculation of the parameters of the noisy channel that connects them (e.g. mean ratio and variance), and by scrutinizing the most prominent sources of errors (we have already mentioned the over-resolution issue). This information could then be used by the scholar when examining automatically tagged texts from outside the pilot corpus.

Currently, the corpus of novels with complete manual annotation that we have at our disposition for this type of evaluation (the pilot corpus, corresponding to the Corpus-Huygens) is still rather limited (the Namescape gold standard corpus not have manually tagged NE resolution).

4. Future work

Apart from finding ways of dealing with noise, we have several other wishes for next steps in this research. Obviously, one would want to optimize 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 and refine, and publish their data, we need to to implement additional functionality. After uploading documents to the NE tagger, the researcher should be able to use the exploration and visualization tools on his/her own data. To be able to deal with the “noise” problem described above, the scholar 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 texts tagged by themselves and (if there are no IPR issues at stake) make them available to other users.

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.

[19] There are many translated titles in this corpus. It would be interesting to compare the outcomes of similar experiments on the original (mostly English) versions. Unfortunately, we do not have the originals at our disposal.

Chinchor, Nancy, Erica Brown, Lisa Ferro, and Patty Robinson (1999), 1999 named entity recognition task definition, *Technical report*, MITRE. http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.pdf.

Cristianini, Nello and John Shawe-Taylor (2000), *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1 ed., Cambridge University Press.

Desmet, B. and V. Hoste (2010), Towards a Balanced Named Entity Corpus for Dutch, *in* Calzolari, N., K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association, Valletta, Malta, pp. 535–541.

Desmet, Bart and Véronique Hoste (2014), Fine-grained Dutch named entity recognition, *Language Resources and Evaluation* **48** (2), pp. 307–343.

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*.

Li, Dingcheng, Karin Kipper-Schuler, and Guergana Savova (2008), Conditional random fields and support vector machines for disorder named entity recognition in clinical texts, *Proceedings of the workshop on current trends in biomedical natural language processing*, Association for Computational Linguistics, pp. 94–95.

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.

Namescape (2013), Namescape demonstrator documentation, *Technical report*.

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_13.

Sutton, Charles and Andrew McCallum (2012), An introduction to conditional random fields, *Foundations and Trends in Machine Learning* **4** (4), pp. 267–373.

Tjong Kim Sang, Erik F. (2002), Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition, *Proceedings of CoNLL-2002*, Taipei, Taiwan, pp. 155–158.

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.

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.