1. Project Title & Acronym and Abstract

Title: Mapping the Landscape of Names in Modern Dutch Literature
Acronym: Namescape

Abstract: Recent research has conclusively proven names in literary works can only put fully into perspective when studied in a wider context (landscape) of names either in the same text or in related material (the onymic landscape or "namescape"). Research on large corpora is needed to gain a better understanding of e.g. what is characteristic for a certain period, genre, author or cultural region. The data necessary for research on this scale simply does not exist yet. The proposed project aims to fill the need by annotating a substantial amount of literary works with a rich tag set, thereby enabling the participating parties to perform their research in more depth than previously possible. Several exploratory visualization tools will help the scholar to answer old questions and uncover many more new ones, which can be addressed using the demonstrator. The main tools will be made available as CLARIN compliant web services for use in other contexts.
Target Start Date: 1 January 2012
Target End Date: 1 January 2013
Type: Demonstrator Project
Call: Open Call

2. Coordinator

Name: dr. Karina van Dalen-Oskam
Function: Research leader Department of Literary Studies / Textual Scholarship
Organization: Huygens ING - KNAW
Address: Prins Willem-Alexanderhof 5, 2595 BE Den Haag
E-mail: karina.van.dalen@huygens.knaw.nl
Tel: 070-3315875
Fax: 070-3820546
Role(s): User, Data Provider

3. Composition of the Project Team

Name: dr. Maarten Marx
Function: Assistant Professor
Organization: University of Amsterdam
Address: Science Park 904, 1098 XG Amsterdam
E-mail: maartenmarx@uva.nl
Tel: 020 525 2888
Fax:
Role(s): Technology Provider

Name: lic. Katrien Depuydt
Function: Leader Taalbank Nederlands
Organization: Institute for Dutch Lexicology (INL)
Address: Matthias de Vrieshof 3, gebouw 1171, 2311 BZ Leiden
E-mail: katrien.depuydt@inl.nl
Tel: 071-5272479
Fax:
Role(s): Technology Provider, Data Provider

4. CLARIN centre

The project results will be made available on a server of the Institute for Dutch lexicology (CLARIN centre).

5. Requested Budget

€ 120.000

6. Description of the Proposed Project

6.1 Research Question(s)

The research questions are: What is the usage and what are the (stylistic and narrative) functions of names in literary texts? Starting out from simple objectives like studying the relative proportion of names versus non-names in a text and the proportion of use of different types of named entities in typical literary works, quantitative overviews of names and name types etc. will be used to draw conclusions about functions of names that could not be highlighted before in a verifiable and repeatable way. A pilot analysis of 32 Dutch novels is done, but needs to be tested and expanded on a much larger corpus. A few examples of questions emerging from earlier work on the pilot corpus serve to illustrate the type of research:

- Earlier work (van Dalen-Oskam, 2005) suggests that the ratio between the use of first names and family names is indicative of the level of intimacy in a novel
- Another interesting element is the distribution of so-called "plot internal" versus "plot external" names. Plot external names refer to persons, places or objects outside the fiction (Obama, Buenos Aires, Lord of the Rings) which seem mostly used as characterizations of plot internal, fictional characters. Different novels and authors etc. make a different use of these plot external names and for different reasons. It would be extremely useful to be able to get a quick overview of the ratio between plot internal and plot external names in a large corpus of novels, to learn more about their possible functions.
- Quantitative study of the pilot corpus led to the identification of two functions of geographical names that have not been noticed before: a higher use of different geographical names embodied a geographical taboo in a novel, and the pronouncing of lists of geographical names functioned as a calming mantra for the main character.

The demonstrator will enable scholars to check these observations in a much larger corpus tagged for named entities than otherwise would have been humanly possible to

analyze. The exploratory function of the visualization tools in the demonstrator are expected to lead to many more new observations, questions, and inspirations.

6.2 Research Data

A corpus of 582 Dutch novels written and published between 1970 and 2009 will be used. Part of this corpus (550 novels) is available to the INL in XML format. The remaining 32 novels (some of which are English novels translated into Dutch) are in ASCII.
All of the material is based on OCR scans of the originals and may contain scanning errors. Metadata (title, year of publication, publisher, page numbers) is available but needs restructuring and completion (e.g. adding with ISBN). The metadata will be made CLARIN compliant (CMDI).
IPR reasons prevent the corpus as a whole or any substantial part of it to be made publicly available. Cf. section 7, D1. Open information repositories, such as Wikipedia, will be used for resolving named entities.

6.3 Technology

INL proposes to use its own adaptation of the Stanford named entity recognizer (Finkel, et al, 2005) for robust handling of OCR errors. It will be trained and customized to handle a more fine-grained annotation of named entities. The tagged names will be mapped to real entities (named entity resolution) using Dutch and English Wikipedia by means of an extension of the technique from (Meij et al 2011).
In a bootstrapping cycle, the texts are tagged by a preliminary NE tagger and results are inspected and corrected in a browser-based annotation tool (already available at INL).
The demonstrator is based on an XML database system which allows for complex content and structure search (the open source eXist DB system) and a website which is built solely using W3C approved XML-languages (Relax NG, XSLT, XQuery). Via the database all books will be coupled to other sources using their ISBN as a key (Google Books, DBNL, Amazon.com, Wikipedia, summary ("uittreksel")-sites).
To quickly inspect the occurrence of entities in one book, in all books by an author, or in a group of books in one genre we will use the *barcode browsing* technique developed at UvA. This depicts a source (a book) as a column of tiny bars, each representing a paragraph. Bars containing an entity are coloured according to the entity type. This technique improves upon the well-known State of the Union visualization created by the NY Times.[1]
The entities in each novel will be modeled as nodes of an undirected weighted network where the weighted edges are given by co-occurrence counts. Each network is made available as a valid GraphML (http://graphml.graphdrawing.org/) XML file. For each entity, a parsimonious language model [Hiemstra et al 2004] is created based on the lemmatized words occurring around the entity. We visualize network and language models together as described in [Kaptein et al 2009] and recently demonstrated in the attackogram of the Algemene Beschouwingen 2011 (http://debat.politiekinzicht.com) (UvA). In case of many entities and relations we first perform hierarchical cluster

[1] http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html and http://xml.politicalmashup.nl/bb/index.xql?q=naam&col=%2Fbb-open%2FEwoudSanders

analysis and present the network as in the Prinsjesdag dictionary (http://www.inl.nl/over-het-inl/wat-doet-het-inl/nieuws/318-prinsjesdagwoordenboek) of the INL.

6.4 Description

The named entity tagging and resolution enables quantitative and repeatable research where previously only guesswork and anecdotal evidence was feasible. The visualisation module will enable researchers with a less technical background to draw conclusions about functions of names in literary work and help them to explore the material in search of more interesting questions (and answers).
Users from other communities (sociolinguistics, sentiment analysis, …) will also benefit from the NE tagged data, especially since the NE recognizer will be available as a web service, enabling researchers to annotate their own research data.

7. Literature

Karina van Dalen-Oskam, 'Altijd naar het noorden. Gebruik en functie van namen in *Boven is het stil* van Gerbrand Bakker'. Rich Internet Publication (in preparation) at http://www.cs.uu.nl/research/projects/i-cult/xposre/demo/huygens/
Karina van Dalen-Oskam. 'Vergleichende literarische Onomastik' in: Brendler, A. und S. Brendler (Hrsg.). *Namenforschung morgen: Ideen, Perspektiven, Visionen.* Hamburg: Baar, 2005, p. 183-191. English translation online, Comparative Literary Onomastics, Dutch version online: http://www.huygens.knaw.nl/wp-content/bestanden/pdf_vandalenoskam_2005_Vergelijkende_literaire_onomastiek.pdf

Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. 2004. Parsimonious language models for information retrieval. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04). ACM, New York, NY, USA, 178-185.

Rianne Kaptein, Maarten Marx, and Jaap Kamps. 2009. Who said what to whom?: capturing the structure of debates. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09). ACM, New York, NY, USA, 831-832.

Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, Maarten de Rijke, Mapping queries to the Linking Open Data cloud: A case study using DBpedia, Web Semantics: Science, Services and Agents on the World Wide Web, Available online 28 April 2011, http://www.sciencedirect.com/science/article/pii/S1570826811000187

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf