Discovering an Encyclopaedic Novel: a case study in automatically analysing Harry Mulisch's *The Discovery of Heaven* (1992)

Leon van Wissen
Research Master student in Dutch literature
l.van.wissen@student.vu.nl

Marieke van Erp
Postdoc in the Computational Lexicology and Terminology Lab
marieke.van.erp@vu.nl

Ben Peperkamp
Professor of Modern Dutch literature
b.j.peperkamp@vu.nl

The literary studies field has a longstanding tradition of detailed analysis of literary works. This results in fine-grained, but usually small-scoped studies. The advent of computational methods makes it possible to scale up the subject of analysis and start for instance comparing entire oeuvres of authors or even genres. Before we do so though, it is important to evaluate the precision and impact of such computational methods, for which we have carried out a small study in which we automatically analysed Harry Mulisch's *The Discovery of Heaven* (1992) using DBpedia Spotlight (Daiber et al., 2013). We chose to investigate this novel as is considered by many as Mulisch's masterpiece (Brems, 2006), it is a fair body of work (nearly 1.000 pages, containing ≈270.000 words) and contains many references to disciplines such as the natural sciences, theology, humanities and politics. The novel embodies and uses encyclopaedic knowledge and could therefore be seen as an encyclopaedic novel. One could say it strives to capture the ideas and opinions of its time into its narrative, and shows a variety of means to interpret the world (Mendelson, 1976). These aspects make this kind of novel ideal to be analysed using computational methods, given the fact that the overwhelming amount of information it grants is hard to be grasped by the novel's reader. (Van Ewijk, 2011, p. 214)

The goal of this study is twofold:

1) Map which scientific disciplines (here seen as a set of practices within scientific communities, regarding domains of research and accepted theories and practices) are represented in Harry Mulisch's *The Discovery of Heaven.*
2) Assess the added value of computational resources and semantic web tools such as DBpedia and DBpedia Spotlight in complementing traditional literary analysis.

We wrote a program that, in common words, takes a (Dutch) text file, scans the file for word combinations that have their referent on Wikipedia, and classifies these words into scientific disciplines. It outputs a list of terms with their discipline and a network graph, visualizing clusters of knowledge domains that are represented in the text. The first step is being done by feeding chunks

of text into DBpedia Spotlight, which automatically filters out so called Named Entities, for example person names such as 'Julius Caesar' or location names such as 'Cuba'. The program then tries to match this entity to one or multiple disciplines found in DBpedia that we have predefined in a list, such as 'Physics' or 'Biology'. This is done by crawling through DBpedia's hierarchical structure of semantic data, by listing every hierarchical parent of a Named Entity's subject. DBpedia is generated from the structured data in Wikipedia such as infoboxes and category information. The program then combines this information into a network graph (Figure 1).

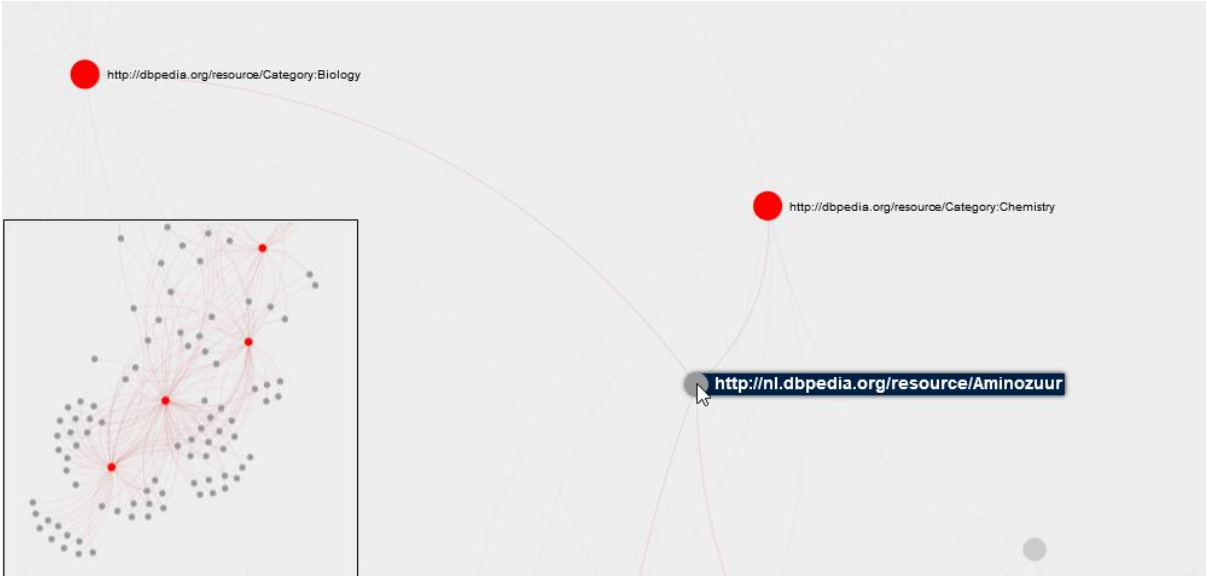

Figure 1: Screenshot showing two disciplines from the first pages (prologue, pp. 7-17) of *The Discovery of Heaven* with corresponding entities linked to them. Highlighted is the entity 'Amino acid' which is linked to both Biology and Chemistry. At the bottom left there is an overview shown of the entire network. Taken from: http://kyoto.let.vu.nl/~vanerp/TheDiscoveryOfHeaven/.

The result shows that a lot of the found entities could be linked to categories of disciplines, which should give insight in the way which knowledge is represented and distributed in the novel. The program is able to mark cross-disciplinary entities found in the text, so that a term like 'amino acid' is both linked to 'Biology' and to 'Chemistry'. Statistical information combined with information from the graph shows which domains are particularly represented in the novel. On this ground the program could be useful as a tool for researchers, used at a first exploration of a novel, paving the way for further close-reading and expanding the corpus from a single novel to a set of encyclopaedic novels. Also, the program offers a repeatable and consistent method to annotate a text. All of this analysis is done in less than it takes to annotate and analyse the novel manually.

Naturally, care should be taken in using these automatic annotations, but the DBpedia Spotlight tool allows for inspection and correction of individual tags. In this experiment, we did not adapt the Spotlight tool to the domain, and some clean-up, for example where it tries to link fictional characters, should ideally be performed. We plan to implement filtering techniques to refine the results and we expect to get much cleaner results when applying those.

One feature of *The Discovery of Heaven* that makes this an interesting use case for further natural language processing research is the fact that translations exist in 16 languages. We are currently working on gaining access to at least some of these to expand our research to a comparison of NLP

tools between the translations. The method we use is not bound to a single language and should in theory work for every text you input in all languages in which there are Wikipedia-articles.

References

Brems, H. (2006). *Altijd weer vogels die nesten beginnen*. Amsterdam: Bakker, pp.556-557.

Daiber, J., Jakob, M., Hokamp, C. and Mendes, P. (2013). Improving Efficiency and Accuracy in Multilingual Entity Extraction. *Proceedings of the 9th International Conference on Semantic Systems*.

Van Ewijk, P. (2011). Encyclopedia, Network, Hypertext, Database: The Continuing Relevance of Encyclopedic Narrative and Encyclopedic Novel as Generic Designations. *Genre*, 44(2), pp.205-222.

Mendelson, E. (1976). Encyclopedic Narrative: From Dante to Pynchon. *MLN*, 91(6), p.1267.

Mulisch, H. (2010). *De ontdekking van de hemel*. 20th ed. Amsterdam: De Bezige Bij.