



## Supervisor's statement of a final thesis

**Student:** Andriy Nazim  
**Supervisor:** Ing. Milan Dojčinovski, Ph.D.  
**Thesis title:** Extraction of linguistic information from Wikipedia  
**Branch of the study:** Web and Software Engineering

**Date:** 3. 6. 2019

<i>Evaluation criterion:</i>	<i>The evaluation scale: 1 to 4.</i>
<b>1. Fulfilment of the assignment</b>	<i>1 = assignment fulfilled, <b>2 = assignment fulfilled with minor objections,</b> 3 = assignment fulfilled with major objections, 4 = assignment not fulfilled</i>
<i>Criteria description:</i> Assess whether the submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.	
<i>Comments:</i> The main tasks of the thesis were to: i) get familiar with the DBpedia NIF dataset, ii) analyse the SOTA for extraction of lexical information, iii) develop a method for extraction of linguistic information based on the DBpedia NIF dataset, apply it on several languages and model the results using Ontolex, and iv) implement a UI for querying the dataset.  The assignment has been fulfilled with two minor objections: 1) the student had to execute a more detailed analysis of the state-of-the-art approaches 2) the student did not store the results in the Ontolex format, nor it discusses the decision made to store the results in a proprietary format.	
<i>Evaluation criterion:</i>	<i>The evaluation scale: 0 to 100 points (grade A to F).</i>
<b>2. Main written part</b>	<b>60 (D)</b>
<i>Criteria description:</i> Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies? Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 26/2017, Art. 3. Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.	

*Comments:*

In general, the thesis contains all expected parts, starting with an introduction and background and related work chapters, followed by a chapter describing the main work/contribution and a chapter describing the evaluation part.

However, the thesis is difficult to read, follow and understand. The main issues with the thesis are as follows:

- Although the student is not an English native speaker, I would expect a better language expressions.
- Some parts require more detailed information
  - DBpedia NIF, which is the core resource used in the thesis, could be better described. Current description, as part of section 1.1.6, is very short and poorly described.
  - The main focus of the thesis is the extraction of linguistic information, however, the type of linguistic information considered in the thesis (i.e. synonyms, homonyms), are poorly described. Section 1.1.1. on Synonyms, homonyms, semantic relationships should provide more details.
- In the related work section, the student has primarily focused on existing linguistic datasets (WordNet, BabelNet, etc.). However, actual approaches for extraction of linguistic information are poorly covered (Section 1.2.5).
- The text providing explanations for particular steps of the approach could be better structured. E.g. on page 25 - it is difficult to map the Steps with the diagram on Figure 21.
- Listings of code for execution of simple tasks are not required - e.g. code listings on page 27, 28, 29
- There are several missing citations:
  - page 6, figure 12 - missing citation for the table
  - page 12, figure 19
  - page 13, figure 13
  - page 27, figure 22
- A more logical title for Chapter 2 is "Background and State-of-the-art", instead of "State-of-the-art" since it also provides background information on the concepts, models and resources used in the thesis.

*Evaluation criterion:*

*The evaluation scale: 0 to 100 points (grade A to F).*

**3. Non-written part, attachments**

70 (C)

*Criteria description:*

Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

*Comments:*

There are three contributions, wrt to the technical implementation:

- 1) set of scripts written for the process of generation of the required datasets - the main contribution of the thesis
- 2) a set of three datasets - synonyms, homonyms and semantic relationships
- 3) simple Web application (UI) for searching and browsing the results

The main problem with the non-written part is that the implementation is not documented (apart from the thesis itself). There are 13 scripts but the order of execution is undocumented. Also, the code is not accompanied with comments.

Another problem is the actual model used to capture the final result - synonyms and homonyms. The initial idea was to use the Ontolex, however the student has decided to define own, simpler but less expressive model. The problem with this model is that it is not possible to match a particular count value with the actual surface form.

*Evaluation criterion:*

*The evaluation scale: 0 to 100 points (grade A to F).*

**4. Evaluation of results, publication outputs and awards**

60 (D)

*Criteria description:*

Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

*Comments:*

The plan is to integrate the results within the DBpedia core dataset. However, this still requires further investigation on the quality of the data.

The results itself are not novel, a similar dataset based on Wikipedia content has been generated by the BabelNet knowledge base authors.

Overall, the work has to be improved, in terms of being configurable and scalable, so that it can be deployed in practice.

*Evaluation criterion:*

*The evaluation scale: 1 to 5.*

**5. Activity and self-reliance of the student**

- 5a:  
1 = excellent activity,  
**2 = very good activity,**  
3 = average activity,  
4 = weaker, but still sufficient activity,  
5 = insufficient activity
- 5b:  
1 = excellent self-reliance,  
2 = very good self-reliance,  
**3 = average self-reliance,**  
4 = weaker, but still sufficient self-reliance,  
5 = insufficient self-reliance.

*Criteria description:*

From your experience with the course of the work on the thesis and its outcome, review the student's activity while working on the thesis, his/her punctuality when meeting the deadlines and whether he/she consulted you as he/she went along and also, whether he/she was well prepared for these consultations (5a). Assess the student's ability to develop independent creative work (5b).

*Comments:*

The student regularly attended the meetings and was always prepared.  
The student required guidance but he managed to independently develop the technology.

*Evaluation criterion:*

*The evaluation scale: 0 to 100 points (grade A to F).*

**6. The overall evaluation**

65 (D)

*Criteria description:*

Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.

*Comments:*

The main goal of the thesis was to derive lexical information from Wikipedia article texts.  
In general, the student fulfilled the assignment and managed to apply knowledge acquired during the studies.  
However, there are few minor issues with this thesis: the language and the content of the thesis could be improved, the student had to pay attention more on the model used for storing the results, and the actual implementation could be configurable and better documented.  
Nevertheless, the above-mentioned problems are not crucial for the actual quality and results of the thesis.

Considering all comments above, I recommend grade D.

Signature of the supervisor: