Digital Tools and Methods in Conceptual History

thomas jurczyk

This presentation was held in the *CERES Computer Café* (November 8, 2019).

General Introduction

This presentation aims at providing an overview of the overall approach of my thesis "The Notion of *surb* in Ancient Armenian Texts from the Fifth Century CE" with a focus on the digital methods and tools used in the examination. The first part of the presentation provides an overview of the content and methodology[1], whereas the second part focusses on the digital tools and methods.

Research Question of my Dissertation

My dissertation "The Notion of *surb* in Ancient Armenian Texts from the Fifth Century CE" deals with two interrelated research questions.

The first research question focusses on a specific semantic problem. What does the Armenian notion *surb*[2], including its derivatives, mean in early Armenian texts from the fifth century CE?

The translation of *surb* as *holy* is implicitly assumed in many contemporary English translations of ancient Armenian texts. This observation leads to the second research question that undertakes a comparative study by relating the word field(s) around the notion *surb* to both the broader modern English notion of *holy* and the more narrow notion of *holy* in the academic field of religious studies. The major aim of the second research question is to look for potential differences as well as overlappings between the three semantic fields. This should serve to evaluate whether these notions can rightfully be seen as interrelated with each other and thus form a semantic field of *holy* across time.

[1] I am still not sure how one would properly distinguish between "methods" and "methodology" in English. In German, one would normally speak of "methods" when talking about the concrete methods used in a dissertation, whereas "methodology" refers to the critical discussion of these methods.

[2] Note that the use of italics indicates a focus on the general notion rather than on the actual word. The general notion might include derivatives, inflected forms, as well as synonyms. For instance, in the case of the general notion *holy*, different words such as sacred, holiness, or sanctified are usually included. When talking about the particular word "holy" as it appears in a text, quotation marks are used.

Figure 1: enter image description here

Corpora

To deal with these research questions, approaches from the field of corpus linguistics are applied. Corpus linguistics ...

> ... is a scientific method of language analysis. It requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language (...) In essence, corpus linguistics is a quantitative methodology; this means that corpus linguistics typically works with numbers which reflect the frequencies of words and phrases in corpora (McEnery & Hardie 2011).[3]

The examinations in my thesis are based on three different text corpora. Two of these corpora were created be me, whereas the third corpus is the already existing *English Web 2015 (enTenTen2015)* corpus.
The first corpus includes Armenian texts and is divided into the *Ancient Armenian Surb Corpus* (AASC) and the *Ancient Armenian Full Text Corpus* (AAFTC). Both corpora are based on the same texts, namely

- the "History of the Armenians" by Agat'angełos,
- the "Life of Mašt'oc"' by Koriwn,

[3]Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge, New York, NY: Cambridge University Press, 2f.

- the "Epic Histories,"
- and the Gospels in their Armenian translation according to the so-called Zohrab-Bible.

The AAFTC includes the full version of these texts (mostly types only)[4], whereas the AASC includes only sentences and passages in which the terminology around *surb* appears. The AASC also includes a lemmatized version of these sentences. The English corpora are divided into two corpora of very different sizes. The first corpus is the *Holy/Sacred English Corpus* (HSEC) including text samples from different genres (academic texts, religious texts, encyclopedia entries, tweets). This corpus as well as the Armenian corpora were built by me. The second English corpus is the already existing *English Web 2015* (*enTenTen2015)* corpus that is used with the help of *Sketch Engine*.

Corpus Linguistics and Corpus Analysis as "Digital" Methods?

Overview

The methodological approach of my thesis is split into a macro- and microanalysis. Digital tools and methods are only applied in the macroanalysis. Thus, the details of the microanalysis will not be discussed in this presentation. In the macroanalysis, I am mainly using approaches from the field of corpus linguistics.[5] Most of them are closely related to ideas related to distributional semantics and can rightfully be described as text statistical methods. Most parts of these analyses on the macro level are done with the freely available software package #LancsBox.[6] For more information on #LancsBox, please see the following sections.
After the corpus analysis, the results are first annotated, then visualized, and eventually compared to one another. The overall "pipeline" of the methodological approach of this thesis is displayed in the following diagram.

"Processing pipeline" {: .image-caption}

In the subsequent parts of this presentation, I will discuss selected steps in this overall approach together with the corresponding digital tools and methods.

Corpus analysis (text statistics)

The methods used in the macroanalyses are mostly related to text statistics. For the purpose of this thesis that is primarily concerned with semantic questions,

[4]For the differences between tokens, types, and lemmas, see this link.

[5]See Rayson, Paul. 2015. "Computational tools and methods for corpus compilation and analysis." In The Cambridge Handbook of English Corpus Linguistics, edited by Douglas Biber, Randi Reppen, Douglas Biber, and Randi Reppen, 32–49. Cambridge: Cambridge University Press.

[6]For an introduction of text statistics in corpus linguistics, see: Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge, New York, NY: Cambridge University Press.

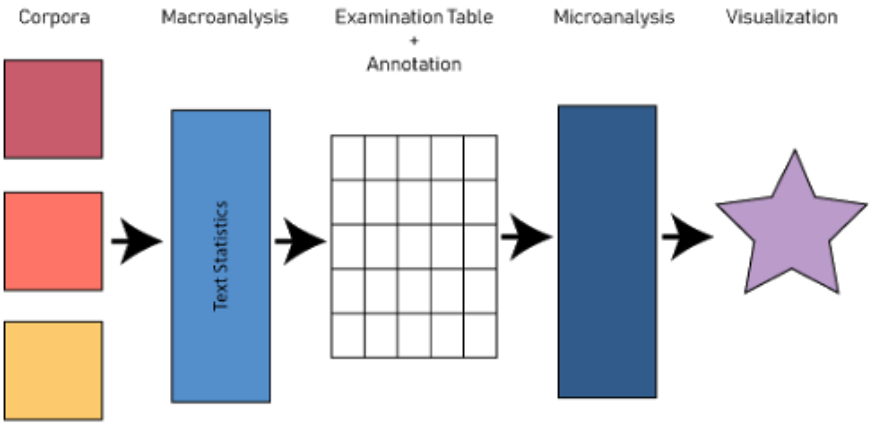

Figure 2: enter image description here

the application of very basic text statistics seemed to be sufficient. The following three text statistical methods were applied in the macroanalysis.

- **Word Frequency Lists**The idea of frequency lists is probably the most straight forward method in order to gain insights into the distribution of lexical units in a text (corpus). Besides the raw word frequencies (called absolute frequency, short *AF*), it is also common to calculate normalized frequencies, for instance, the relative frequency, short *RF*.

- **Analysis of Keywords**The analysis of keywords is based on the comparison of word frequency lists between a corpus of interest and a reference corpus.
  > A corpus of interest (C), sometimes referred to as a 'focus corpus' (Kilgariff 2012) or 'node corpus' (Scott 1997), is compared with a baseline reference corpus (R) using a statistical measure to identify words that are used either more often or less often in C when compared to R.[7]

- **Collocation Analysis** The collocation analysis is in the center of the examination in this thesis as it precisely tries to find out which other lexical units (and thus semantic fields) appear commonly in the context of words such as "surb" and "holy." According to the ideas behind distributional semantics, neighboring words of a lexical unit (such as "holy") can help to derive the contextual meaning of the lexical unit. Thus, looking at the words that most commonly appear in the context of certain notions in a text or corpus is a great way of getting information of their overall meaning that derives from the sum of their contextual use cases.

[7]Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge, New York, NY: Cambridge University Press, 80.

The association measure used in this thesis is *log-likelihood* (LL). For the math behind *log-likelihood* (LL), see (Brezina 2018: 72). *Log-Likelihood* (LL) is a regularly used association measure in corpus linguistics that is preferred in this thesis instead of the several other existing measures due to its somewhat non-exclusive character that also highlights the frequency of collocations. The collocation window is usually L2-R2; however, there might be cases where L3-R3 is preferred.

Annotation

The results of the corpus analysis are then grouped in a table for each text and annotated based on a twofold annotation system. The first part of the annotation scheme has been developed by me ("domain" in the table below). The second scheme ("semantic_field" in the table below) is based on an alread existing annotation system called *UCREL Semantic Analysis System* (short USAS) developed and maintained at Lancaster University. One row in these examination tables represents one word (lemma) that appeared in, at least, one part of the macroanalysis.

lemma	keyword	word_freq	coll_freq	sum	domain	semantic_field
religion	1	1;200	3;43	98	concept	religion_S9

The values in the "sum" column are based on the results of the macroanalysis and represent the overall importance of the lemma compared to other lemmas in the examination table.

Visualization

The examination tables of the texts and the "sum" values in particular are then taken to examine the application and meaning of the terminology in the texts. Furthermore, the examination tables are compared to one another to look for possible differences as well as overlappings in the contextual use of the respective notions. Both the examination of the single tables as well as their comparison are done with the help of visualized graphs.
Two examples of such graphs are shown below. The first graph displays the distribution of "domains" in the AAFTC/AASC "Gospels." The second graph visualizes the overlappings between words in the examination tables of the AAFTC/AASC and the HSEC. Both visualizations have been created with Gephi using the *Fruchterman-Reingold* algorithm.

AAFTC/AASC "Gospels" domain graph including words. {: .image-caption}

Overlappings between "semantic fields" of AAFTC/AASC and HSEC (particularly good quality ^_^) {: .image-caption}

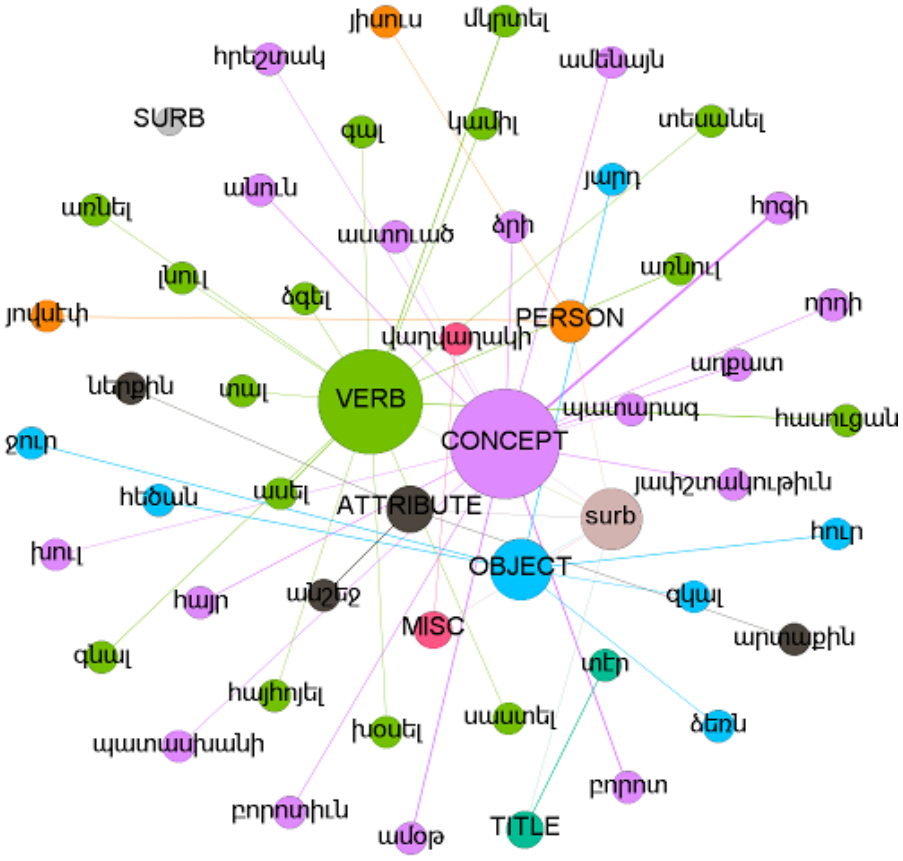

Figure 3: Graph 1: Distribution of “domains” in AAFTC/AASC Gospels

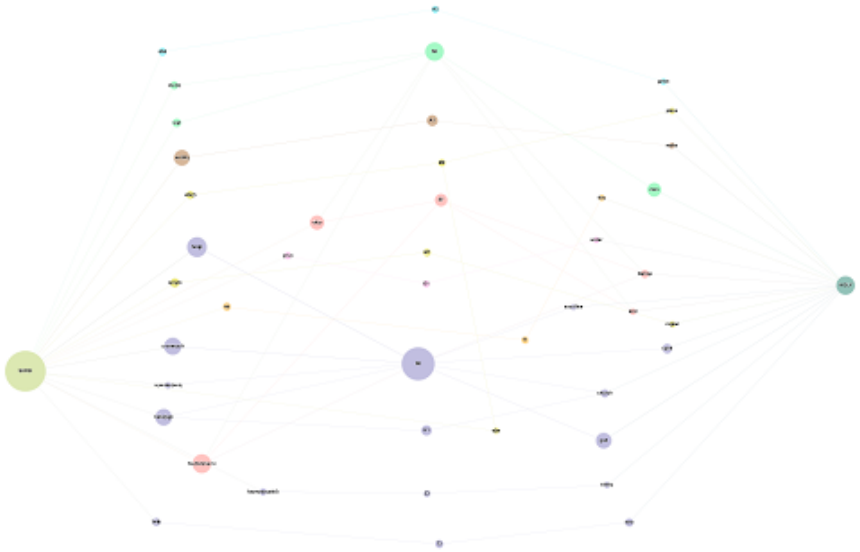

Figure 4: Graph 2: Overlappings between “semantic fields” of the AAFTC/AASC and HSEC.

Digital Tools and Methods

The second part of this presentation focusses on selected examples of digital tools and methods used in my thesis. The overall idea is not to give a comprehensive overview of all the parts in which certain digital tools or methods are applied. Instead, I would like to present selected examples that best illustrate potentially useful ways of applying digital tools and methods in a dissertation that is focussed on conceptual history.
Therefore, I have divided the following overview into the discussion of ...

1. ... ready-to-use software packages for text statistics and visualization.
2. ... high level programming languages such as Python and some powerful libraries that can be used for corpus analysis.

"Ready-to-use" software solutions

There are mainly three ready-to-use software packages that I applied in several parts of my thesis.

#LancsBox

The first program is #LancsBox[8], a freely available software package developed by, among others, Vaclav Brezina and Tony McEnery at Lancaster University. According to the self-description on their website, #LancsBox is ...

> ... a new-generation software package for the analysis of language data and corpora developed at Lancaster University

Together with the already-mentioned book "Statistics in Corpus Linguistics,"[9] #LancsBox provides great tools and features to ...

- ... either load in your own corpus or already existing corpora (for instance, the *British National Corpus* BNC).
- ... preprocess your texts with the help of features such as lemmatization or part-of-speech tagging (for English texts only).
- ... conduct many text statistical analyses such as the ones applied in my thesis.
- ... visualize the results of the respective analyses.
- ... save the results in different file formats.

Gephi

The next tool that I would like to mention is Gephi, an "open graph viz platform" that is, according to the description on their website, ...

[8] Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), 139-173. Online available here.

[9] See footnote 4.

> ... the leading visualizaton and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

According to wikipedia.org, a graph is defined in the following way.

> In mathematics, and more specifically in graph theory, a graph is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related". The objects correspond to mathematical abstractions called vertices (also called nodes or points) and each of the related pairs of vertices is called an edge (also called link or line). (Trudeau, Richard J. (1993). Introduction to Graph Theory (Corrected, enlarged republication. ed.). New York: Dover Pub. p. 19.)[10]

Gephi offers great features to visualize and work with graphs by applying different graph algorithms. Yet, in order to properly analyze and visualize the graphs, the data already needs to have a certain structure. Thus, preproccessing the graph data is usually a necessary first step before using Gephi. Graphs are mainly used in the last part of my dissertation in order to display and compare the relations between the words in the examination tables.

Sketch Engine

Sketch Engine is an online portal that is, at least for RUB members, free to use. According to the, slightly advertising, self-description on the corresponding website, Sketch Engine defines itself in the following way.

> Sketch Engine is the ultimate tool to explore how language works. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage. It is also designed for text analysis or text mining applications. (...) Sketch Engine contains 500 ready-to-use corpora in 90+ languages, each having a size of up to 30 billion words to provide a truly representative sample of language

The major reason for the integration of Sketch Engine is its inclusion of different large-sized text corpora. In my thesis, I am mainly using the English Web corpus 2015 (enTenTen15) with Sketch Engine that includes 15 billion words in texts collected from the Internet.
What I particularly like about Sketch Engine is its Word Sketch feature. According to the documentation, the Word Sketch feature can be used as a ...

> ... one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc. In my thesis, I am using the Word Sketch feature as an

[10] Graph (discrete mathematics) on wikipedia.org

opportunity to compare its results to the results of the analysis of the small to medium-sized corpora. It is thus primarily applied to verify or falsify the results of the analyses of the self-built corpora.

Python

Besides the ready-to-use software packages, I am writing small scripts or programs in the high level programming language Python for different purposes in several parts of my thesis. The use of programming languages such as Python enables the creation of much more flexible and specific solutions compared to the use of ready-to-use software packages that are often much more restrictive. However, for complex problems, it is usually preferred to use already existing software solutions instead of trying to program everything from scratch.
A good example are the text statistical methods used in this thesis. In theory, most of the statistical methods could also be programmed from scratch. Yet, using existing software solutions does not only save a lot of time and is generally more convenient, but it also makes it less likely to get false results due to insufficient testing or wrong implementations of the algorithms. On the other hand, using programming languages makes it particularly easy to create "processing pipelines" that integrate different steps in a more holistic approach, for instance, by automating data transformations. In addition, coding also allows the flexible implementation of individual features that might not be part of existing software solutions.

Collecting, processing, and storing data

A major part of the use of Python in my dissertation is ocncerned with collecting, processing (cleaning), and storing data.

A good example of the use of Python in collecting data is the creation of the previously mentioned text corpora. Particularly the HSEC data was mainly taken from already digitized text sources that are available online. Among these texts were encyclopedia entries that could simply be copy & pasted. Yet, other subcorpora such as the HSEC "Twitter Data" needed a more specific approach. In the case of the Twitter data, I wrote a Twitter streamer with the help of the Python library Tweepy to set up a automated script on my website on Pythonanywhere.com. This script collected and still collects tweets from the live Twitter stream that include certain keywords such as "holy." The major tweet sample used in my thesis consists of approximately 12,000 tweets. The current sample (6th of November 2019) includes approximately 32,000 tweets.
Another example of the use of Python in the process of collecting data is the lemmatization of the AASC. In this case, I set up a web scraper with Python that made use of the already existing lemmatization of the Armenian version of the Bible on arak29.com. The remaining lemmas that were not part of the Bible data were then manually added by me. A full manual lemmatization of the AASC texts would have taken much more time. With a web scraper, the

collection of the lemmas and the creation of a respective dictionary as well as the lemmatization of the Armenian sentences only takes a couple of seconds.

The step of (pre-)processing the raw data (Tweets, digitized books, texts collected via web scraping, etc.) consists of, at least, three different steps including ...

1. ... cleaning,
2. annotating.
3. transforming/storing the data.

Particularly in the case of data cleaning, some basic programming skills and a basic understanding of regular expressions often turn out to be very helpful. For instance, the HSEC "Twitter Data" needed to be cleaned of smilies and hyperlinks, whereas the HSEC "Canon Law" data included many paragraph symbols and other structural elements that would have impeded the later corpus analysis ("noise"). In these cases, finding a fitting software solution for each cleaning task or even doing a manual cleaning of the texts would either be very difficult or very time consuming (or both). Yet, with only some very basic knowledge of programming languages such as Python, these tasks can be solved rather quickly.

The storing of the data includes, among others, setting up rudimentary databases such as SQLite as well as the transformation of data from one file type or format into other file formats or types. This might, at first, sound trivial,[11] but depending on the indivual case, the transformation and preparation of data stored in a specific format into another format can become tricky without having the flexibility of programming languages.

Analysing data with Python (libraries)

The last aspect that I would like to mention is the enormous potential of Python libraries such as *pandas* or *scikit-learn* machine learning with Python.

The *pandas* library is an ...

> ... open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.[12]

This is not the place to go into the details of this very powerful and essential Python library - at least for data analysis purposes. In sum, it can be characterized as a powerful statistical tool that provides extensive possibilities to load in, analyse, and transform different types of data, among them textual data.

The second important library that I would like to mention is *scikit-learn* machine learning with Python. *Scikit-learn* is an, according to the self-description on the website, "simple and efficient tool for data mining and data analysis."[13]

[11] "Why can't I simply use Miccrosoft Office's"Export" feature?!"
[12] pandas website
[13] scikit-learn website

In my thesis, I trained some of the machine learning classifiers such as *KNN* or *Stochastic Gradient Descent* with *scikit-learn* in order to classify the tweets in the HSEC "Twitter Data" into either "religious" or "non-religious" tweets. This worked relatively well and enabled me to add another layer to my analysis by separating the religious from the non-religious use cases of words related to *holy*. The topic of machine learning is yet another complex topic that should be discussed in a separate session of the *CERES Computer Café*. However, the application of basic machine learning classifiers is relatively easy and offers interesting new perspectives on the material and possibilities for future research. For a concise introduction into machine learning with Python, I recommend the tutorial series by *sentdex* on YouTube[14] as well as the book "Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: Concepts, tools, and techniques to build intelligent systems."[15] Yet, both require an at least rudimentary understanding of Python in order to follow along.

Conclusion

I hope that I have been able to share some impressions and experiences of how and where to use digital tools and methods in a dissertation concerned with the conceptual history of notions such as *holy*. This presentation was not meant to provide detailed insights into the actual examination itself or its results. Instead, the idea was to demonstrate the overall methodology and work flow as well as different digital tools with their respective applications in my thesis. I hope to have shown that the use of digital tools and methods can offer new and unconventional perspectives on established research questions. Furthermore, it facilitate different tasks that would otherwise be relatively teadious to accomplish, among them the gathering, cleaning, and storing of data.

Footnotes

---

[14] sentdex' Channel on YouTube

[15] Géron, Aurélien (2019): Hands-on machine learning with Scikit-Learn, Keras and TensorFlow. Concepts, tools, and techniques to build intelligent systems. 2nd ed.