Attempts at Visualisation of Etymological Information

Armin Hoenen
Goethe University Frankfurt
Juridicum, Senckenberganlage 29,
hoenen@em.uni-frankfurt.de

Abstract content

1. Introduction

Reconstructing word histories constitutes an important part of lexicographers (especially etymologists) work. We would like to present a rather simple and then a more complex hypothesis for a word history exemplarily, both along with concurrent visualizations. The key question is how to derive useful visual representations of the histories of single words representing the content of articles from etymological lexica.

2. Study Object and Related Work

Our main objects of study are single words, the histories of which we would like to trace. It must be said, that in etymological print lexica, visualizations are no mainstream phenomenon. One reason may be that drawing and printing visualizations can be relatively cumbersome (in comparison to text) given the print medium. Furthermore, a textual representation was required in any case. With the advent of the digital and especially effective automatic extraction, conversion and visualization methods, the question of adding value by visualization comes into focus. The only work explicitly focusing on this issue known to the author is Dixit and Karrfelt (2016) who use Etymological WordNet by De Melo (2014) as basis for their visualization. While visualization for etymological relations seems understudied, in recent years with the large scale migration of content from print to digital representation and the emergence of primarily digital resources, data on etymology has been transported into the digital medium. A large lexical resource in this respect is the DWDS, see Klein and Geyken (2010), which comprises under more digitized versions of several large German lexica among which the ”Etymologisches Wörterbuch” and the lexicon of the Grimm brothers which contain many etymologically relevant articles. For the simple visualization attempt, we will use data from this resource. Just as articles in the Wikipedia have been produced in a primary written form, such resources are mainly textual in content. Wikipedia quite soon has become the object of intense study in Computer Science and especially the Linked Open Data community has spent a lot of effort to extract information from the Wikipedia in a structured way and derive various knowledge bases from it, the most famous project being the DBPedia from Auer et al. (2007). The same has happened to a much smaller extent for etymological textual data. De Melo (2014) and Sagot (2017) use Wiktionary as their basis for the extraction of etymological patterns, whereas Chiarcos and Sukhareva (2014) and Abromeit et al. (2016) use more specialized data and explicitly etymological dictionaries such as the Turkic Etymological Dictionary. Consequently information can be interpreted as modelled as a graph, where words or morphemes typically form nodes and are connected by relations or typed edges. Typically those types carry labels such as ”derived from”, ”cognate”, ”variant orthography” or ”etymological origin”. Can visualization of such graphs help grasp etymological relationships more effectively than when forced to read and evaluate longer textual representations?

2.1. Visualization as Added Value

All but sign languages have a very sequential character that is one word has to follow the other, see also Ong (2013). This implies that textual representations are at a loss when it comes to presenting multidimensional relations, a phenomenon remedied only partly by interlinked hypertext. Especially for the representation of etymological word histories, which are often complex involving many languages and alternative hypotheses, a good visualization could become a means of effectively transporting this information, more effectively so than pure text. If this effectiveness is achieved, then visualization can be used to save effort and time and increase comprehension.

3. A Simple (?) Case

As a real world example for a simple case, we use the German word ”Schatulle” - casket (also small ornate box for valuables). Can we generate visualizations of such etymological relations on a larger scale and relatively easily add them to digital representations? Partly, this is being done. The informational foundation has already been laid, see De Melo (2014). For new data, the way of conduct would be an extraction of such patterns from the text, which as in the DBPedia may be tricky at times and may lead to some loss without further fine-tuning, compare Abromeit et al. (2016). A result could look like Figure 1.

3.1. A Complex Case

The history of Japanese SHA-KAI (社会) ’society’ free after Yanabu, Akira (柳父 章 (Yanabu, 1991). Initial problem: Such a word does not exist in the early Meiji-era (starting 1868) in Japan, absence of a translation equivalent and missing awareness of any semantically equivalent entity in contemporary Japanese society. Society has 2 main extant senses, see the OED.[1] One with a local implication naming a group of people such as the National Geographic Society, the other relating to the larger context of all individuals of

[1] http://www.oed.com/view/Entry/183776?redirectedFrom=society

e.g. a state. Sketch of the word history, free summarization after (Yanabu, 1991) with additional references:

- the two constitutive ingredients:
  - SHA – originally, the Chinese character 社 (in its modern Japanese pronunciation SHA) was referring to the shrines of earth gods (as opposed to for instance air gods 神 KAMI and villagers ceremonial meetings around them as 社会 SHAKAI, see also Matsumura (2006)
  - KAI – more traditionally, this Chinese character 会 refers to meeting, but also to the fit of something, to some (harmonic) togetherness
- Early translations: friends (NAKAMA), meet (ATSUMARU), government (SEIFU) . . . in the Ekaijiten dictionary (1847/48): KAI (Meeting), kessha (group; SHA expresses group of people sharing the same goals)
- Since sense 2 was largely missing for translations of Western works, some new terms were coined: NINGEN KOUSAI, from NINGEN (mankind) and KOUSAI (typically delimited human relationships - master and servant etc.)
- In the broad public meanwhile groups form which discuss Western cultural artifacts (and texts). They call themelves something-SHA. The probably most important SHA is the MeirokuSHA, which issues the journal Meirokuzasshi concerned with new Western phenomena.
- SHAKAI and KAISHA both are attested as generalizations when talking about the phenomenon of those groups. KAISHA: (any) -SHA: head = SHA, KAI = meet SHAKAI: rather the phenomenon of meeting in such SHAs, (attested: SHAKAI ENZETSU), head = KAI
- SHAKAI forms new bonds in the lexicon and becomes some antonym of SEKEN simple folk, ordinary people

At this point, translators presumably start using SHAKAI for society, sense 2, on a large scale and the word enters common vocabulary.

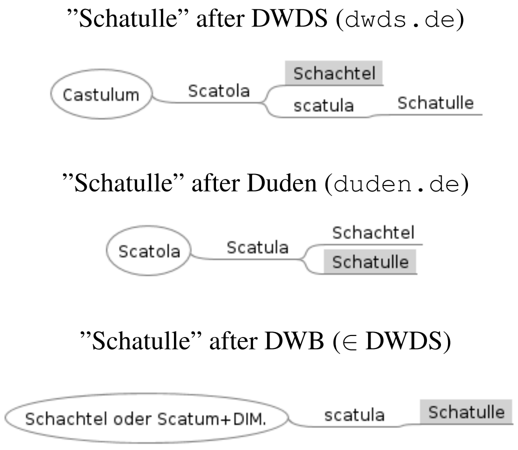
Alternatively, parallel to English, leaving *National Geographic* out, *Society* remains. For Japanese however, the single character SHA – because of a) it is mostly used as affix where a usage as a single word – not as an affix – would be perceived as unusual and because b) its singulary reading (many characters if isolated as a word have to be read in a different way than if in compounds) of YASHIRO which means temple/shrine is dominant – may have not been linguistically fit for this purpose. A loanword SOSAITI is attested, but would not manifest, probably because the contemporary need to coin a new generalized expression for the something-SHAs temporally (and in subject) coincides with the arrival of a new semantic concept, society sense 2, which needed to be named. Although the visualization in Figure 2 is by no means more than a clumsy attempt, for its generation a large variety of very different semantic and temporal relations had to be integrated. A simple graph based visualization may not be the best and most effective visualization to be found for complex word histories.

4. Discussion

It is evident that cases cannot be classified binarily into simple and complex very easily. The distinction is technically inspired and should refer or be mapped to a threshold of complexity (thus a binary decision boundary) where for cases which display more complexity, a simple visualization would be too error-prone. Guided by this principle, the threshold should be chosen rather too low than too high. There are many possible ways of potential measurement of this kind of complexity with factors such as article length, the number of matched relations in the article, the number of cross-referenced dictionary entries, the depth or breadth of the concurrent graph and so forth. An empirical study could provide better insights.

5. Conclusion

We think that for rather simple word histories, effective visualizations are possible and possibly extractable while for more complex cases many more layers of information and more complex visualization have to be considered so that at the current point in time not only machine readable complex data, but also structured models for their effective visualization are largely absent. In the presentation, I will therefore display attempts at the visualization of simple and complex cases and finally try to extend visualizations to attempts at the visualization of more complex etymological processes (for instance the introduction and discussion of many new terms in Japans Meiji era and the concurrent transformation of the word network) or processes whereby older words get used more and more pejoratively (comp. Dornseiff and Waag (1955)).

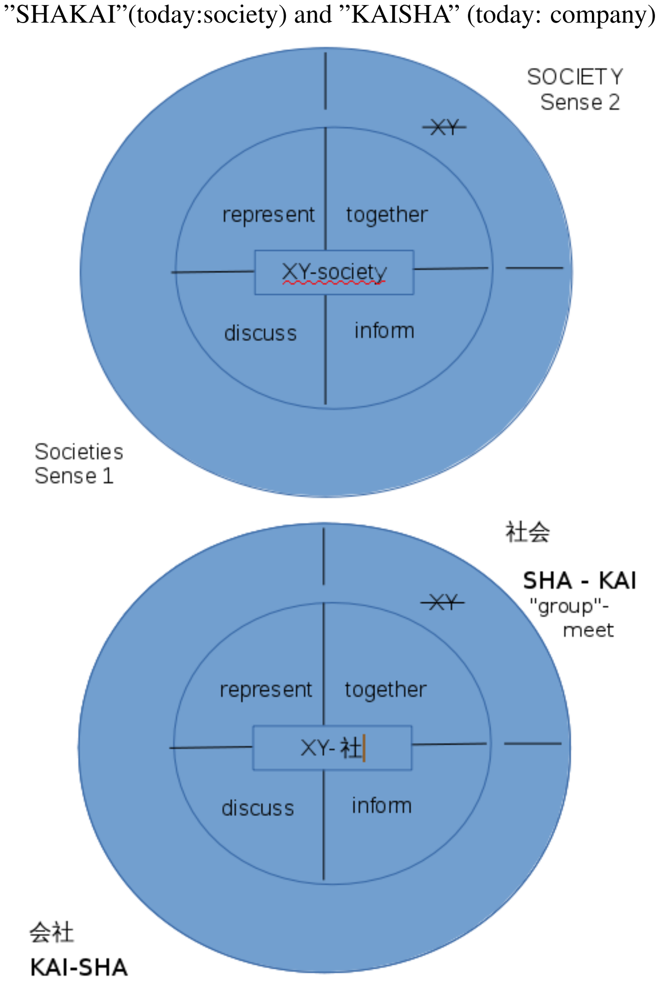


”Schatulle” after DWDS (`dwds.de`)

”Schatulle” after Duden (`duden.de`)

”Schatulle” after DWB (∈ DWDS)

Figure 1: Visualizations of different hypotheses on the etymology of German *Schatulle*. Note, that the third visualization is displaying two alternative hypothesis represented in one node. The hierarchical tree structure, which is also common to many other etymological resources allows in this case the use of the mindmapping software Freemind.



”SHAKAI”(today:society) and ”KAISHA” (today: company)

Figure 2: Visualizations attempt of the more complex word history of SHAKAI - society. The outer circle depicts generalization. Above: English, Below: Japanese.

6. Bibliographical References

Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016). Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources, Portoroz, Slovenia*, pages 11–19.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735.

Chiarcos, C. and Sukhareva, M. (2014). Linking Etymological Databases. A Case Study in Germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41.

De Melo, G. (2014). Etymological wordnet: Tracing the history of words. In *LREC*, pages 1148–1154.

Dixit, C. and Karrfelt, F. (2016). Visualizing etymology: A radial graph displaying derivations and origins.

Dornseiff, F. and Waag, A. (1955). *Bezeichnungswandel unseres Wortschatzes: ein Blick in das Seelenleben der Sprechenden*. Schauenburg.

Klein, W. and Geyken, A. (2010). Das digitale wörterbuch der deutschen sprache (dwds). In *Lexicographica: International annual for lexicography*, pages 79–96. De Gruyter.

Matsumura, A. (2006). 2006. *Super Daijirin*, 1.

Ong, W. J. (2013). *Orality and literacy*. Routledge.

Sagot, B. (2017). Extracting an etymological database from wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, pages 716–728.

Yanabu, A. (1991). *Modernisierung der Sprache: eine kulturhistorische Studie über westliche Begriffe im japanischen Wortschatz*. Iudicium-Verlag.