EXTENDED ABSTRACT

Rule-based search in historical text databases - Visualization techniques

Thomas Pilz, Wolfram Luther, Universität Duisburg-Essen
{pilz,luther}@inf.uni-due.de

The project Rule-Based Search in Historical Databases with Non-Standard Spellings (RSNSR, Pilz et al. 2005) will provide an online-available search-engine that can be used by interested amateurs as well as professional linguists. Parallel to the implementation of a customizable software architecture to support an efficient search functionally recalling all relevant historical spellings of a modern word in the underlying search text, there is an ongoing research focusing on the construction of rule bases and distance measures (Kempken et al. 2006) in reference to historical epochs or isoglosses. In a recent paper (Kempken et al. 2007) we describe several tree map techniques used to visualize among other aspects the productivity of rule sets in deriving non-standard spellings in old German texts.

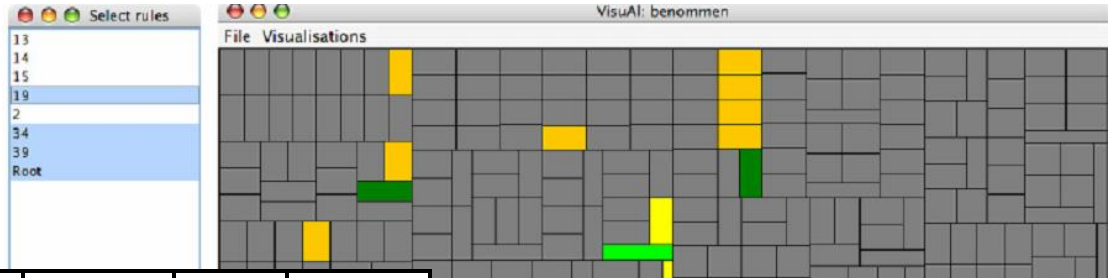
According to the derivation process a rectangular area is recursively subdivided into a set of smaller rectangles alternating between vertical and horizontal subdivision. Each rectangle represents a node of the tree and the enclosed sub-rectangles correspond to all descendants of this node. Similar visualization techniques help finding diachronically and diatopically representative replacement sequences (Bruls et al. 2000; Fekete and Plaisant 2002; Shneiderman 2006). A recently conducted study proves that treemaps ease the understanding of rule hierarchies, the detection of productive and non productive rules and the evaluation of a rule's importance. They also provide better search performance. In a further publication it was shown that the approach is not restricted to the German language (Archer et al. 2006)

The paper (Kempken et al. 2006) determines requirements for an adequate visualization tool. It should allow the detection of relevant rule sequences. A sequence of rules is considered relevant if it leads to a historical variant appearing in the text base (Established Spelling). Irrelevant sequences that lead to non-existing variants should be pointed out in parallel.

The tool should make it easy to find permutations of rules that produce the same spellings, discern patterns to describe characteristics of non standard orthography (depending on location and period). It should allow for deriving upper bounds for the length of relevant rule sequences and for providing a means of accessing extensive amounts of spellings.

Twelve subjects participated in a development application study, doing particular tasks using treemaps with appropriate color schemes, tool tips, and control panels to switch certain rules on and off as well as a functionality to zoom into the tree. The evaluation of the treemap approach was carried out in opposition to a text-based display of the derivation process. With the results obtained from the answers we claim the following advantages for our treemap solution to the rule visualization problem:

- A huge amount of data and information about the derivation process is conveniently presented in one screen.
- Our approach allows for parallel processing of all rule-based derivations in one picture. The user receives a quick overview of all derivations and the rules involved.
- A serial perspective on the process is achieved using the zoom option; text-based information corresponding to a certain derivation step is presented on demand via tool tips.
- Special tasks can be achieved like the identification of derivations that most affect the whole derivation process or the determination of the minimum set of required rules to derive the established spellings in text.



Color/ Meaning	Established spelling	Selected Rules	Redundant rule set
Gray	No	No	No
White	Yes	No	No
Yellow	No	Yes	No
Light green	Yes	Yes	No
Orange	No	No	Yes
Dark green	Yes	No	Yes

Figure 1: Treemaps show established and irrelevant spelling variants

However, it seems necessary to deepen the study with more sophisticated treemap visualization techniques (Fekete, Plaisant 2002) and a larger group of subjects to obtain more significant results on the benefit of the treemap approach with respect to the user's performance.

Another interesting approach is to visualize established spelling variants and meta-data of corresponding texts. The following figure shows authors of a certain epoch and corresponding spelling variants in their work. It would be of interest to refine the starlike plot into a distance graph structure that shows similarity and proximity.

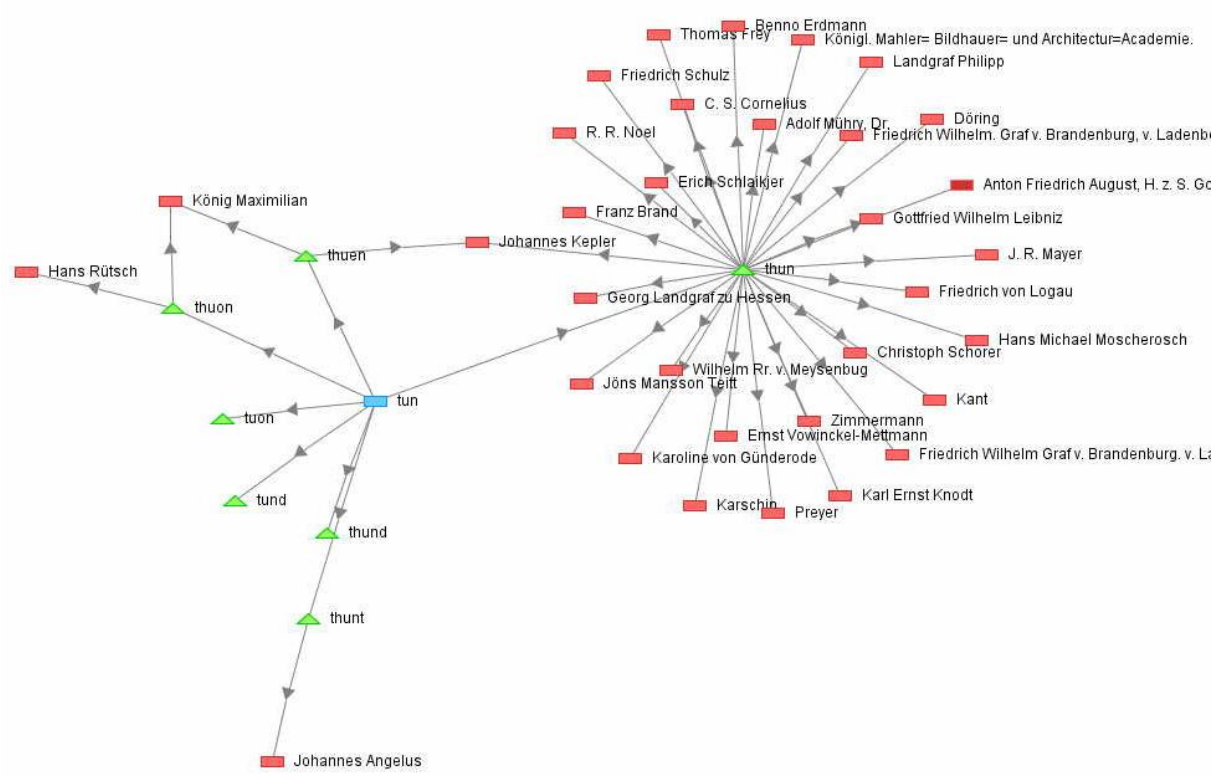


Figure 2: Graph representation of our training database

A further goal of the RSNSR project is the development of temporal and local filters for phonetic rules and the revision of rules through consideration of the text base and statistical analyses of the occurrence of spelling variants. A prototypical realization of an interface supports the interactive visualization of language variation (Schmidt 2006) allows the definition of several parameters to show isoglosses running between different regions of Germany and to cluster text samples of different epochs and their writings.

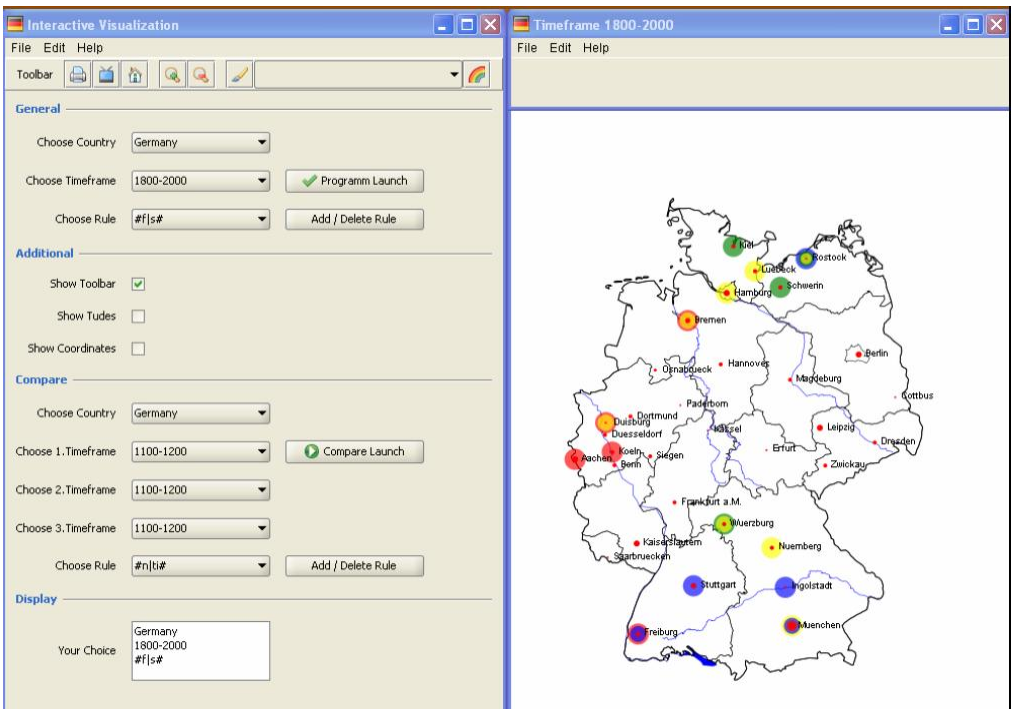


Figure 3: User interface and rule base allocation depending on time, region and text epoch

In concluding, we will emphasize that visualization facilities are an important part within a more general Literature Research System to support the user in accessing and working with historical texts (Biella et al. 2005).

References

D. Archer, A. Ernst-Gerlach, S. Kempken, T. Pilz, and P. Rayson: The identification of spelling variants in English and German historical texts: Manual or automatic, In Proceedings of the Digital Humanities Conference, pp.3 –5, 2006.

D. Biella, E. Dyllong, W. Luther & Th. Pilz: An On-line Literature Research System with Rule-Based Search, In Proceedings of the 4th European Conference on e-Learning (ECEL2005), 10-11 November 2005, Amsterdam, Netherlands, ISBN 1-905305-12-5, pp. 67 – 76.

D. Bruls, C. Huizing, and J. van Wijk: Squarified treemaps, In Proceedings of the joint Eurographics and IEEE TCVG Symposium on Visualization, pp. 33 – 42, 2000.

J. Fekete and C. Plaisant: Interactive information visualization of a million items, In Proceedings of the IEEE Symposium on Information Visualization 2002 (InfoVis), pp. 117 – 124, 2002.

S. Kempken, W. Luther, and T. Pilz: Comparison of distance measures for historical spelling variants, in Proceedings of the IFIP AI Conference, pp. 295 – 304, 2006.

S. Kempken, W. Luther, Th. Pilz: Visualization of rule productivity in deriving non-standard spellings, To appear in Proceedings SPIE 2007.

T. Pilz, W. Luther, U. Ammon, and N. Fuhr: Rule-based search in text databases with nonstandard orthography, In Proceedings ACH/ALLC, 2005.

J.- D. Schmdt: Interaktive Visualisierung von raum-zeitbasierten Textsorten und ihren Relationen. Diploma thesis, Duisburg 2006.

B. Shneiderman: Treemaps for space-constrained visualization of hierarchies, 2006.
http://www.cs.umd.edu/hcil/treemaphistory/