Visualizing the evaluation of distance measures

Thomas Pilz

University of Duisburg-Essen
Faculty of Engineering
Department of Computer Science
pilz@inf.uni-due.de

Axel Philipsenburg

University of Duisburg-Essen
axel.philipsenburg@uni-due.de

Wolfram Luther

University of Duisburg-Essen
Faculty of Engineering
Department of Computer Science
luther@inf.uni-due.de

Abstract

This paper describes the development and use of an interface for visually evaluating distance measures. The combination of multidimensional scaling plots, histograms and tables allows for different stages of overview and detail. The interdisciplinary project Rule-based search in text databases with nonstandard orthography develops a fuzzy full text search engine and uses distance measures for historical text document retrieval. This engine should provide easier text access for experts as well as interested amateurs.

1 Introduction

In recent years interest in historical digitization projects has markedly increased, bearing witness to a growing desire to preserve cultural heritage through new media. All over Europe projects are arising digitizing not only monetary but also intellectually valuable text documents. While more and more documents are being digitized and often provided with well designed interfaces, they are not necessarily easy to work with, especially for nonlinguists. Spelling variants, faulty character recognition (OCR) and typing errors hamper if not circumvent sensible utilization of the data. One such example is the archive of Jewish periodicals in German language, Compact Memory (www.compactmemory.de). Even though of great cultural value and very well maintained, the operators of this project simply did not have the resources required to postprocess or annotate their automatically recognized text documents. A user for example searching for the word “Fruchtbarkeit” (=fertility) will not be able to find a certain periodical from 1904 even though it clearly contains the word. Worse, he will not even come to know that this text was missed. Because the full text aligned with the graphical representation of the text contains recognition errors, only the search for the misspelled word “Piuchtbaikeit” instead of “Fruchtbarkeit” finds the correct page (cf. Figure 1). The same problem arises when dealing with historical spelling variation. German texts prior to 1901 often contain historical spelling variants. Numerous projects are dealing with similar problems of optical character recognition or spelling variation.

To meet those problems linguistics and computer science are closing ranks. Fuzzy full-text search functions provide access to nonstandard text databases. Since the amount of data on the one hand and the divergence of users on the other increases day by day, search methods are continuously presented with new challenges. The project RSNSR (Rule-based search in text databases with

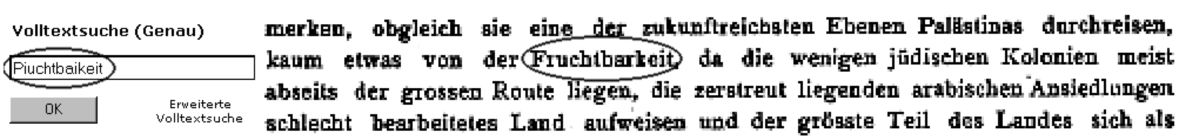

Figure 1. OCR errors prevent successful retrieval on digitized texts if misspelled variants are used for full text search.

Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, pages 84–92,

nonstandard orthography) seeks to improve the retrieval of nonstandard texts. Such texts might include historical documents, texts with regional/dialectal or phonetic variation, typos or OCR errors. The project's funding by the Deutsche Forschungsgemeinschaft (DFG [German Research Foundation]) was recently extended by two years.

2 Comparing similarity measures

One of the important issues in building a search engine for nonstandard spellings is a reliable way to allow the comparison of words, that is, to measure the similarity between the search expression and the results provided. Given the abundance of distance measures and edit-distances available, methods are needed for efficiently comparing different similarity measures. In (Kempken et al. 2006) we evaluated 13 different measures with the calculation of precision and recall to determine which were most qualified to deal with historical German spelling variants. We mainly used our own database of historical spellings, manually collected from the German text archives Bibliotheca Augustana, documentArchiv.de and Digitales Archiv Hessen-Darmstadt. Currently our database consists of 12,687 modern-historical word pairs (that we call *evidences*) originating between 1293 and 1919.

The algorithm that proved best for calculating the edit costs between the modern and the historical spellings is called *Stochastic distance* (SM) and was originally proposed in 1975 by Bahl and Jelinek. In 1997 Ristad and Yianilos (Ristad et al, 1997) took it up again and extended the approach to machine learning abilities. Due to the complexity of language, apparently similar scopes can obviously favor totally different mechanisms. The Variant Detector VARD developed by Rayson et al. to detect spelling variants in historical English texts uses the standard Soundex algorithm with convincing efficiency (Rayson et al, 2005). The same algorithm yields an error rate 6.7 times higher than the stochastic distance for the comparison of German spelling variants. Cases like these suggest that finding one "most suitable" distance measure for all data might not be possible. As soon as the inherent structures change, another measure can prove to be more efficient. Even though, with the SM, we already found a suitable measure, its dependency on the underlying training data forces us to evaluate the training results: what is the size of an optimal training set? Is the training set well chosen? Does 14th-century data appropriately represent 13th-century spellings? Answers to these and similar questions not only help to ensure better retrieval but can also give an insight into phonetic or graphematic changes of language. Since standard calculations of retrieval quality, as we did for the 13 measures, require not only extensive work but are also difficult to evaluate, we propose possibilities for visual evaluation means to speed up and ease this process. The prototype we developed is but one example for those possibilities and is meant to encourage scientists to benefit from visual information representation.

3 Development and functions of an interactive visual interface

Since our project already deals with different methods for calculating word distance, the definition of a generic interface was necessary. Priority was given to the development of a slim and easily accessible device that allows the connection of arbitrary concepts of word distance. Our SM, a rule based measure using regular expressions, Soundex (Knuth, 1973), Jaro(-Winkler) (Jaro, 1995) and a number of additional measures are already implemented in our system. It was built in Java and is embedded in our general environment for the processing of nonstandard spellings.

Information Visualization is a fairly new field of research that is rapidly evolving. A well established definition of information visualization is "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card et al, 1999). While planning the prototype, we also kept Shneiderman's paradigm in mind: "Overview first, zoom and filter details on demand" (Shneiderman, 1996). In dealing with distance measures, our main task is to represent word distance. We employed multidimensional scaling (MDS) to display abstract distance in 2D space (see below). Interactivity is gained with the ability to select and remove spellings from the calculations, lower or raise cutoff frequencies and filters and even change replacement costs with instantaneous effect (see below). This led to a user interface separated into three main views:

- The **Histogram** allows an overview of thousands of data items. The selection of a certain portion of data triggers MDS and table views.

- **Multidimensional Scaling (MDS)** functions as a detail view. Such visualization is used to display sets of several dozen to a few hundred items.
- The **Table View** can display different levels of detail. In (Kempken et al, 2007) we presented a TreeMap approach, another way to display details of single word derivations as an add-on for table views.

3.1 Histograms

Histograms are a widely spread tool for display of statistical distribution of values. In favor of Shneiderman's paradigm, the histogram view represents a combination of overview and zoom functionality. This first stage allows for the reduction of the data set from up to several thousand items down to much more manageable sizes.

To get a first impression of how a spelling distance performs on a set of evidences, we calculate the distance between a spelling variant and the entries in a dictionary. It is ensured that the collection also contains the standard spelling related to the variant. The results are sorted in ascending order by their distance from the spelling variant. Afterwards, the rank of the corresponding spellings is determined. In the best case, the correct relation will appear as the first entry in this list, that is, at the smallest distance from the variant. Often, other spellings appear "closer" to the variant and thus have a higher rank, pushing the spelling we sought for further down the list (cf. Figure 2).

By applying this procedure to a collection of word pairs, we get a distribution of spelling ranks over the set of evidences based on the spelling collection. Good distance measures produce a histogram with most of its largest bars close to the first rank on the left. A good example is the evaluation in section 5 (cf. Figure 5).

	lieb	liebe	lebt
lebt	1.211	1.542	0.0
leib	0.728	1.060	1.243
leibt	1.301	1.632	0.676
lieb	0.0	0.331	1.243
liebe	0.397	0.0	1.641
liebd	0.903	0.991	1.246

Figure 2. The standard spelling "liebe" corresponding to variant "liebd" was pushed back by "lieb" because deletion of <d> is cheaper than the replacement of <d> with <e>.

The histogram provides a good representation of the overall performance of a spelling distance given for a set of test data. The user will quickly notice if a large number of spellings are found in the acceptable ranking range, if there are noticeable isolated outliers or if the values are spread widely over the whole interval. In addition, histograms can be useful as tools for comparing different spelling distances. Usually multiple histograms are viewed one after another or arranged next to each other. While this might be enough to perceive considerable differences in distributions, small-scale variations may pass unnoticed. An easy solution to this problem is to arrange the different histograms in a combined display area, where the relevant subinterval bars are lined up next to one another and made distinguishable by color or texture. Through this simple rearrangement, even small changes become noticeable to the user. Slight height differences between bars of the same value interval can be noticed as can shifts in peaks along the value range.

For more quantitative performance measurement mean value and standard deviation are calculated and presented in numerical form. A distance definition that performs well will have a low mean value as more spellings are found with a good ranking. However, a mean value that is not especially high or low by itself is usually not enough to characterize a distribution. For this reason, it is important to know the values' spread around the distribution's mean value measured by the standard deviation (SD). A distribution with only a few, tightly packed value peaks provides a small SD whereas a widely spread one will have a large SD. A spelling distance that performs well can be recognized by a low mean value accompanied by a low SD. Both key values can also be made visible in a histogram by drawing markers in its background. In this way, even the key values are easy to compare when comparing spelling distances.

3.2 Multidimensional scaling

The MDS view displays smaller subsets, thus allowing further refinement while providing additional information detail.

MDS is a class of statistical methods that has its roots in psychological research. The main application of such techniques is to assign the elements of an item set to a spatial configuration in such a way that it represents the elements' relationships with as little distortion as possible. In this context, MDS can be used to arrange spellings in a two-dimensional space according to their spelling distances from one another. Every available dimension reduces the need for distortion but increases the difficulty to interpret. Two or three dimensions are a good trade-off. This allows for an intuitive display of distances and clusters of spelling variants. It also makes it possible to discover distance anomalies. If this representation is provided with filtering features, it can be used to select subsets of elements quickly and comfortably. These subsets can then be displayed in detailed information views that would be too cluttered with greater numbers of items.

The "distortion" is evaluated by comparing the distances calculated by the spelling distances with the configuration's geometric distances (i.e. distances following geometric rules). A common calculation for this distortion is the so-called "raw stress" factor. Kruskal (Kruskal, 1964) defined raw stress as the sum of distance errors over a configuration. To calculate this error, we use the distance matrix *D*, where each entry holds the calculated distance δ_{ij} between the spellings of the relevant row and column. These values can be modified by $f(\delta_{ij})=a\ \delta_{ij}$ to achieve a scaling more fit for visual distances, thus reducing stress. Comparison with geometric distances also requires this matrix to be symmetric. Because spelling distances are not necessarily symmetric (distance A – B differs from B – A), we use the mean value of both distance directions to create symmetry, as Kruskal suggests. The second part of the error calculation requires the geometric distances d_{ij} between the spellings, which is determined by *i* and *j* of the current configuration *X*. The actual error is the difference between the two distances squared.

$$e_{ij} = \left[f(\delta_{ij}) - d_{ij}(X) \right]^2$$

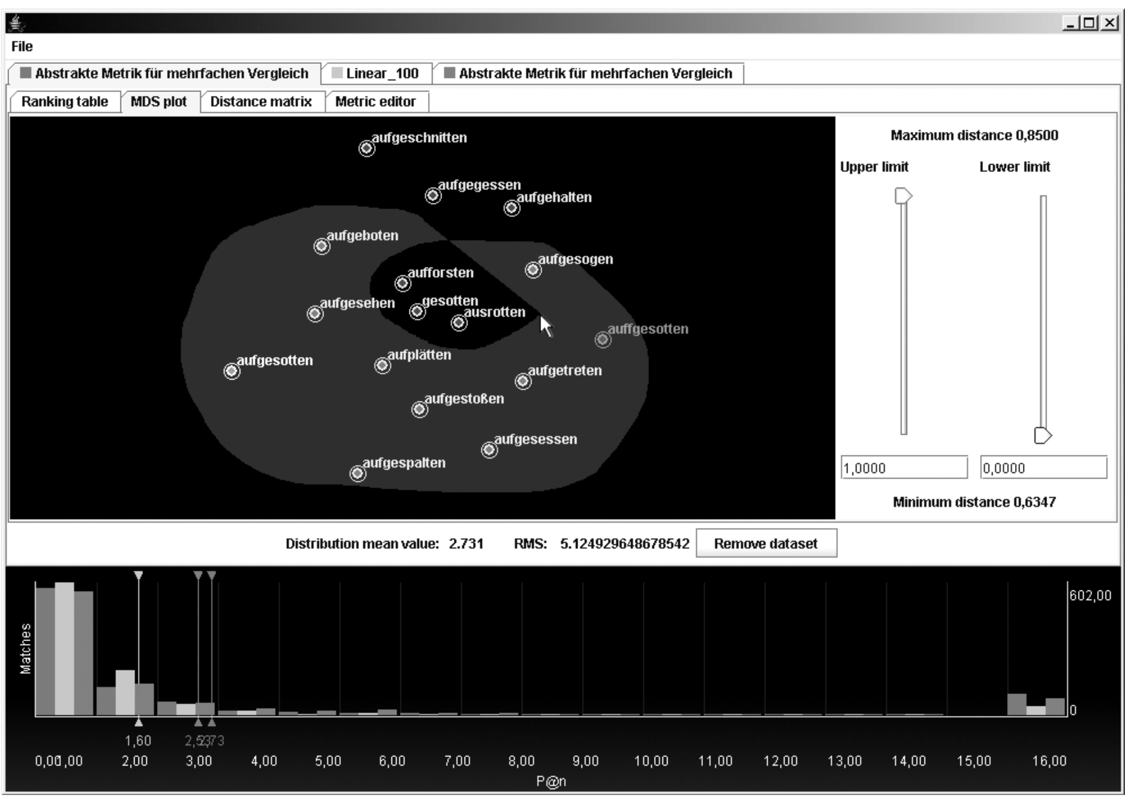

Figure 3. The user interface of the Metric Evaluation Tool showing the evaluation of six metrics trained on different historical training sets, polygon selection in the MDS view and cut-off sliders.

Kruskal's "raw stress" value is then determined by summarizing the error over the elements of the upper triangular matrix. The sum can be restricted to this reduced element set due to the symmetric nature of the matrix.

$$\sigma_r(X) = \sum_{(i<j)} \left[f(\delta_{ij}) - d_{ij}(X) \right]^2$$

In our Metric Evaluation Tool (MET) we used the SMACOF algorithm (see below) to calculate a stress-minimizing configuration. Finding such a configuration is a numerical optimization problem. Because a direct solution of such a problem is often not feasible, numerous iterative algorithms have been developed to calculate an approximate solution close enough to the direct solution, where one actually exists. The SMACOF algorithm (scaling by majorizing a complicated function) is such an approach (De Leeuw, 1977). We start by arranging the items in a checkerboard grid configuration. The algorithm then calculates the raw stress, modifies the current configuration so that it yields a lesser stress value by applying a Guttman Transformation (Guttman, 1968) and then compares the new configuration's stress with the old one. This step is repeated until the change in stress drops below a set threshold or a maximum number of iteration steps is exceeded.

The resulting configuration is usually not an optimal one. Optimal in this case would be a distortionless representation with vanishing stress value. Such a configuration is rarely, if ever, achieved in MDS. There are three main reasons for this:

- Some calculated spelling distances can conflict such that there is no spatial configuration that represents the distances without distortion. For example, a spelling may be determined to be close to several other spellings, which, however, are widely spread out. This is due to the fact that spelling distances do not always fulfill the triangle inequality.
- Although geometric distances, being mathematical metrics, require the spelling distances to be symmetric, the spelling distances calculated will not necessarily be so. For instance, the distance between spellings A and B could be different from that between spellings B and A.
- Even if an optimal configuration were to exist, the iterative optimization process might not actually find it. The algorithm might terminate due to iteration limits or because of being "trapped" in a local minimum.

This restriction on the MDS result, however, is not severe enough to derogate its usage as a visualization tool. Its task is not to reconstruct the calculated distance perfectly but to uncover characteristics of the spelling distances and spelling sets used. These characteristics, such as clusters and outliers, usually outweigh the distortions. Applied to a set of spellings and their distance measure, MDS generates a spatial configuration fit for a plot view. The spellings' positions in relation to one another represent their similarity. Clusters of closely related spellings and outliers are easy to recognize and can be used as starting points for detailed analyses of subsets.

An advantage of this type of visualization is that it considers the calculated distances among all spellings instead of only two. An initial comparison of the difference or similarity of multiple spellings is possible at a single glance and without switching between different views. Additional visual hints can improve the overview even further. Certain spellings, such as the standard spelling or the variant, can be made easily recognizable through color or shape indications. The selection of subsets is aided by zoom and filtering features applied to the plot view. Densely packed clusters can be made less cluttered by changing the plot's zoom factor or by blending irrelevant items into the background. Selecting the spellings by either clicking or encircling allows the subsets to be determined easily. The reduced item set can then be used for a detail view, for example the display of operations and distances like the tabular view. In the MET, the components used to calculate a distance for a given subset can be viewed. In this way, it is easy to understand, for example, why a certain spelling is not as "close" to another spelling as expected.

This visualization approach is applicable to a wide variety of spelling distances as long as they provide a quantitative measurement of two spellings. There are no assumptions made about the distance value except that small values represent a high degree of similarity.

	kundt>kind	kundt>kund	kundt>kunde	kundt>kunz
Distance sum	1.572	0.572	0.903	1.326
del(t) : 0.572	0.572	0.572	0.572	0.572
ins(e) : 0.331			0.331	
repl(d, z) : 0.754				0.754
repl(u, i) : 1.0	1.0			

Figure 4. Table view of replacement costs mirroring deletion, insertion and replacement costs. These costs can be manually adjusted to trigger an MDS view update.

3.3 Tabular views

After refining the selections from several thousand down to a few items, a detailed display of relevant information about the spellings and their calculated distances is needed. At this stage the actual values are more important than a visually comprehensible display of relations.

Two different views in the MET use a tabular arrangement of values. One represents the distance matrix between a set of spellings, similar to the one used to calculate the MDS solution. However, in this case, the distances are not combined to a mean value for both directions. At this point the difference between the two directions can be of interest and should be visible. Standard spelling and spelling variant are displayed in different colors so they can be found more easily.

The second tabular view displays the distances between the standard spelling and the ranked variants. To obtain a better understanding, the results are split up into their components using a Levenshtein-based distance mirroring the replacement costs that occurred when transforming one spelling into the other. These components are displayed in the rows according to their classification, while the different spelling variants appear in the columns (cf. Figure 4). By reordering the columns, the user can move the spellings next to each other in order to compare them more closely.

Another benefit of representing the values in this way is that detailed modifications to the spelling distance can be made interactively. Here, the replacement costs can be changed inside the table itself, allowing an instant evaluation on what effect such a change will have on the distance measure.

4 Interaction

There are several ways to interact with the application. Selection of data triggers an update of the view(s) on the next level of detail: by selecting columns of the histogram, the ranking table is activated; selecting spellings in the ranking table triggers the MDS view where spellings can be selected to be shown in the distance matrix and metric editor. While selections in the tabular views and the histogram can easily be performed with a rectangular selection box, the MDS needed a more elaborate way of selecting data. A polygonal form can be drawn with the mouse that also allows inverted selection (cf. Figure 3). Using two sliders or numerical input, the upper and lower cut-off for selection can be defined. For example, all spellings with a distance higher than 2.5 to the search term can be excluded (cf. right side of Figure 3). Zooming can be performed using the mouse wheel. In the metric editor, showing the highest degree of detail, the costs for the operations of deletion, insertion and replacement can be adjusted. These changes are instantly represented in the MDS view, therefore allowing for the manual calibration of the distance measures (cf. Figure 4).

5 Exemplary application of the interface

To give an example of our MET, we will apply it to a situation we have encountered more than once in the last two years of our research: a set of historical German text documents *T* from between 1500 and 1600 which contains nonstandard spellings. As shown in (Kempken, 2006), the number of spelling variants in old documents is monotonically nondecreasing with advancing age. *T* might also contain errors originating from bad OCR or obsolete characters. Nonetheless, we want to be able to perform retrieval on the document. To simulate a successful full-text search, we manually collected all 1,165 spelling variants *V* in *T* and aligned them with their equivalent standard spellings *S*. We will call those word pairs *evidences*. *S* is now merged into a contemporary dictionary—the OpenOffice German dictionary, which contains approximately 80,000 words. For a reliable evaluation we need a high quality dictionary without typos or historical spellings. The OO-dictionary is the best such wordlist available to us. Our algorithm is able to process dictionaries of up to ~5 million words. Bigger dictionaries can be kept in a database instead of the computer's main memory.

We used the MET applied with six different distance measures to determine the one that works best in finding all the standard spellings *S* "hidden" in the dictionary related to the spelling variants *V*. A normal search task in a historical database would be to find a spelling variant by querying a standard spelling. Because a coherent wordlist of historical spellings was not available, to ensure a more reliable result, we performed the task the other way around. This conforms to the way automatic annotators like VARD work (see above).

Such experiments can be used not only to find the best metric but also to answer general questions:

- Will an SM specifically trained on data from the same time period as *T* work best or will the extension of the time period lower or raise the retrieval quality?
- Is there a level where a "saturation" of training data is reached and the measures' quality cannot be enhanced any further?
- Does the amount of necessary training data vary with the time/location of *T*?

For our first experiment the six measures M_1, $M_2 \ldots M_6$ were trained by the same number of evidences from 14th- to 19th-century German texts. Prior to the training, the evidences had been diachronically clustered (1300-1500, 1300-1700, 1300-1900, 1500-1700, 1500-1900, 1700-1900) into sets, each containing 1,500 word pairs. In general, performance is measured in precision (proportion of retrieved and relevant documents to all documents retrieved) and recall (proportion of retrieved and relevant documents to all relevant documents). Since we ensured that for every historical spelling there is a standard spelling, retrieved and relevant documents are equal and so are precision and recall. We therefore use precision at *n* (*P@n*). This measure is often used in cases where instead of boolean retrieval a ranking of documents is returned, for example in web-retrieval. Precision at 10 is the precision that relevant documents are retrieved within the 10 documents with the highest ranking. In standard settings the MET is using $n \leq 15$.

The task of our prototype now was

- to determine the metric most suitable for the retrieval task, and
- to figure out deficiencies in the metrics to further enhance their quality.

	DMV	SD
1300–1500	*1.37*	*3.174*
1300–1700	*1.384*	*3.222*
1300–1900	*1.261*	*2.983*
1500–1700	*1.375*	*3.1825*
1500–1900	*1.29*	*3.052*
1700–1900	*1.43*	*3.342*

Table 1. Distribution mean value and standard deviation of the evaluated measures

Looking at *P@1* the measures 1300-1500 (58.6%), 1300-1700 (58.7%), 1500-1700 (59.1%) and 1700-1900 (59.4%) seem to be more or less equally efficient. However, by looking at Table 1 we can see that this assumption is not totally correct. The measure trained on evidences from 1700 to 1900 holds a slightly higher distribution mean value and standard deviation than the other two. Interestingly the 1500-1700 measure is not the most efficient one. 1300-1900 and 1500-1900 show better results in *P@1*, DMV and SD. Even though the inclusion of 1300-1500 evidences seems to be of minor significance, the 1300-1900 measure is still slightly better (60.5% *P@1*). Those results are – of course – not significant because of the small dictionary we used. We hope to acquire a bigger freely available dictionary for more expressive results.

The ranking table is now able to show the actual words that led to the result, therefore supporting the expert in further interpretations. The MDS plot and distance matrix let the user explore the words at each rank interactively. Especially interesting are, of course, those words that could not be found within the top 15 ranks. The 1500-1900 and 1700-1900 measures have some difficulties with elder spellings (e.g. *sammatin* [=velvety]). It is also evident that many of the 3.9% of words > *P@10* share certain characteristics:

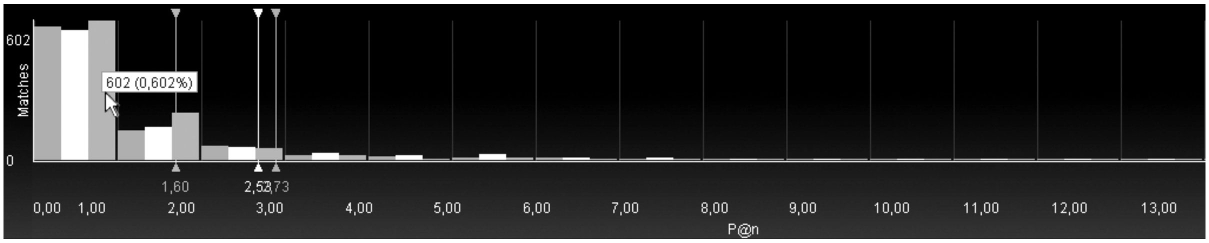

Figure 5. Histogram and DMV comparison of Jaro metric, standard bigram measure and SM 1300-1900.

- a lot of words are short in length (e.g. *vmb, nit, het, eer*). Even a single letter replacement changes a high percentage of the word's recognizability
- some words consist of very frequent graphemes, therefore increasing the space of potential matches in standard spelling (e.g. *hendlen – enden, handeln, hehlen …*)
- some evidences feature high variability (e.g. *ewig – eehefig*)

Those cases complicate successful retrieval.

Comparing the replacement costs in the metric editor (cf. Figure 4) indicates where the SM needs improvement. In our example we noticed that the costs for the replacement of <s> with the German ess-tset <ß> were a little too high, and therefore spellings were not optimally retrieved. A slight manual correction, a control in the MDS view and a recalculation of the histogram showed improved quality of the SM.

Further experiments suggested a "training saturation" (see above) of about 4,000 variants. We trained M_1 on 1,500 evidences from 1300-1900, M_2 on 4,000, M_3 on 6,000 and M_4 on 12,000. While M_1 still shows a small drop in retrieval quality, the differences between M_2 to M_4 are almost unnoticeable. We also performed a cross-language evaluation between historical English and German as we already did manually in (Archer et al, 2006). Our prior results could be confirmed using the MET.

For the comparison of truly different distance measures, as we did in (Kempken, 2006), we used the same data as above with our SM 1300-1900, Jaro metric (Jaro, 1995) and a standard bigram measure (cf. Figure 5). The histogram values of p@<4 for the SM (86.6%) are already 9.2% better than Jaro (77.4%) and 9.9% better than the bigram measure (76.7%). DMV and SD also show how much better the SM performed (cf. Table 2).

	DMV	SD
SM 1300-1900	*1.604*	*3.73*
Jaro	*2.731*	*5.124*
Bigrams	*2.533*	*4.754*

Table 2. DMV and SD comparison of SM, Jaro-Winkler and bigram measure.

6 Conclusion and outlook

While table views will probably not become obsolete any time soon, there are multiple ways to ease and enhance the understanding of abstract data. It has already been documented that users often prefer visual data representations when dealing with complex problems (Kempken, 2007).

In this paper we presented the prototype of our Metric Evaluation Tool and showed that this software is helpful in the evaluation of distance measures. The combination of overview, details and interactivity eases the complex task of determining quality problem-specific distance measures.

Because the MET is a prototype, there is room for improvement. The graphical MDS display could be extended in various ways to further improve the configuration found. Displaying the numerical distance values between spellings as a tooltip or graphical overlay, group highlighting and interactive insertion or removal of additional spelling variants are just a few examples. The bar charts of the histogram view could easily be extended using pixel-matrix displays as proposed by (Hao et al, 2007) to conveniently represent additional information like the distribution of distance ranges.

The MET is only one of the visualization tools we are working on at the moment. No single application will be able to satisfy all the many and various needs that arise in the field of language research. It is our goal to build applications that access and reflect spelling variation in a more natural and intuitive manner. To narrow the field of potentially suitable distance measures, we are also working on automatic text classification. The Word-

Explorer, for instance, is an additional approach to presenting details. Similar to the MDS view in appearance, it is used to further examine words' possible spelling variants, the graphematic space of solution (Neef, 2005). Based on the renowned Prefuse-package for Java (prefuse.org), it provides methods that support easy access and usability, including fisheye, zoom and context menus.

7 Acknowledgements

We would like to thank the Deutsche Forschungsgemeinschaft for supporting this research and our anonymous reviewers whose detailed and helpful reports have helped us to improve this paper.

References

Archer D, Ernst-Gerlach A, Pilz T, Rayson P (2006). The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic?. Proceedings Digital Humanities 2006, July 5-9 2006, Paris, France

Card S K, Mackinlay J D, Shneiderman B (1999). Readings in Information Visualization; Using Vision to think. Morgan Kaufman, Los Altos, California

De Leeuw J (1977). Applications of convex analysis to multidimensional scaling

Guttman L (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points, Psychometrika

Hao M C, Dayal U, Keim D, Schreck T (2007). A visual analysis of multi-attribute data using pixel matrix displays. Proceedings Visualization and Data Analysis (EI 108), Jan 29-30 2007, San Jose, California

Jaro M A (1995) Probabilistic linkage of large public health data file. In: Statistics in Medicine 14, pp. 491-498

Kempken S, Luther W, Pilz T (2006). Comparison of distance measures for historical spelling variants. Proceedings IFIP AI 2006, Sep 8-12 2006, Santiago, Chile

Kempken S, Pilz T, Luther W (2007). Visualization of rule productivity in deriving nonstandard spellings. Proceedings Visualization and Data Analysis (EI 108), Jan 29-30 2007, San Jose, California

Knuth D (1973). The Art Of Computer Programming. vol 3: Sorting and Searching, Addison-Wesley, pp. 391-392

Kruskal J B (1964). Multidimensional scaling by goodness-of-fit to a nonmetric hypothesis, Psychometrika, 29:1-27

Neef M (2005). Die Graphematik des Deutschen. Niemeyer, Tübingen, Germany

Rayson P, Archer D, Smith N (2005). VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. Proceedings of Corpus Linguistics 2005, July 14-17 2005, Birmingham, UK.

Ristad E; Yianilos P (1997). Learning string edit distance. Proceedings of the Fourteenth International Conference, July 8-11 1997, San Francisco, California

Shneiderman B (1996). The eyes have it: A task by data type taxonomy for information visualization. Proceedings Symposium of Visual Languages, Sep 3-6 1996, Boulder, Colorado