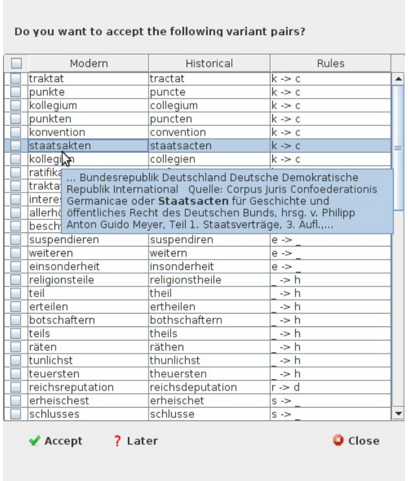
Interactive Rule-generation for Searching in Historic Documents

Andrea Ernst-Gerlach, Dennis Korbar, Ara Awakian

Non-standard spelling produces problems when searching in historic parts of digital libraries. Most users will enter search terms in their contemporary language which differs from the historic language in the documents. Thus many documents can not be found. We overcome this disadvantage with a rule-based approach [Ernst-Gerlach/Fuhr 06] in order to be able to cover the complete vocabulary (and thus increase recall).

Due to the dependency of rules on time and region, rule sets have to be generated when suitable corpora become available. Previous approaches (e. g. [Hauser et al. 07], [Pilz 09] and [Baron/Rayson 08]) need at least in the beginning a lot of manual interaction. We present the *RuleGenerator* user interface that offers interactive support to the user.

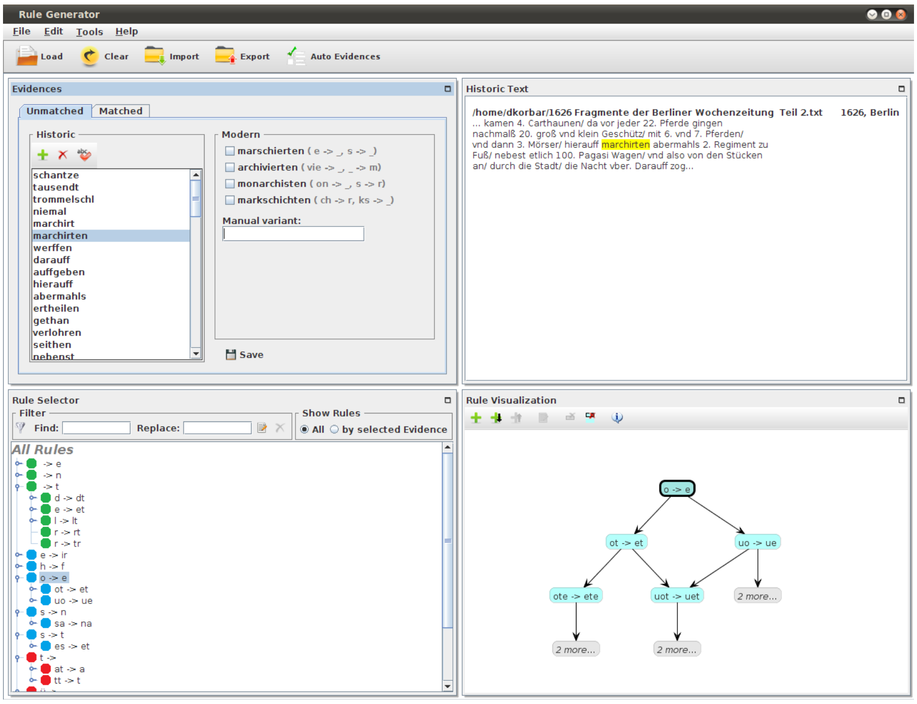


Do you want to accept the following variant pairs?

□	Modern	Historical	Rules
□	traktat	tractat	k -> c
□	punkte	puncte	k -> c
□	kollegium	collegium	k -> c
□	punkten	puncten	k -> c
□	konvention	convention	k -> c
□	staatsakten	staatsacten	k -> c
□	kollegien	collegien	k -> c
□	ratifika		
□	trakta		
□	intere		
□	allerhö		
□	besch		
□	suspendieren	suspendiren	e -> _
□	weiteren	weitern	e -> _
□	einsonderheit	insonderheit	e -> _
□	religionsteile	religionstheile	_ -> h
□	teil	theil	_ -> h
□	erteilen	ertheilen	_ -> h
□	botschaftern	bothschaftern	_ -> h
□	teils	theils	_ -> h
□	räten	räthen	_ -> h
□	tunlichst	thunlichst	_ -> h
□	teuersten	theuersten	_ -> h
□	reichsreputation	reichsdeputation	r -> d
□	erheischest	erheischet	s -> _
□	schlusses	schlusse	s -> _

... Bundesrepublik Deutschland Deutsche Demokratische Republik International Quelle: Corpus Juris Confoederationis Germanicae oder **Staatsacten** für Geschichte und öffentliches Recht des Deutschen Bunds, hrsg. v. Philipp Anton Guido Meyer, Teil 1. Staatsverträge, 3. Aufl.,...

Figure 1: Automatic Evidences



schantze
tausendt
trommelschl
niemal
marchirt
marchirten
werffen
darauff
auffgeben
hierauff
abermahls
ertheilen
gethan
verlohren
seithen
nebenst

□ marschierten (e -> _, s -> _)
□ archivierten (vie -> _, _ -> m)
□ monarchisten (on -> _, s -> r)
□ markschichten (ch -> r, ks -> _)

/home/dkorbar/1626 Fragmente der Berliner Wochenzeitung Teil 2.txt 1626, Berlin
... kamen 4. Carthaunen/ da vor jeder 22. Pferde gingen
nachmalß 20. groß vnd klein Geschütz/ mit 6. vnd 7. Pferden/
vnd dann 3. Mörser/ hierauff marchirten abermahls 2. Regiment zu
Fuß/ nebest etlich 100. Pagasi Wagen/ vnd also von den Stücken
an/ durch die Stadt/ die Nacht vber. Darauff zog...

```
-> e
-> n
-> t
   d -> dt
   e -> et
   l -> lt
   r -> rt
   r -> tr
e -> ir
h -> f
o -> e
   ot -> et
   uo -> ue
s -> n
   sa -> na
s -> t
   es -> et
t ->
   at -> a
   tt -> t
```

Figure 2: RuleGenerator application

The *RuleGenerator* [Awakian 2010] (see figure 2) is divided horizontally in the components *SmartEvidencer* (top) for building the training set and *Rule-Modification* (bottom) [Korbar 2010] for the rule construction. The *SmartEvidencer* contains two components:

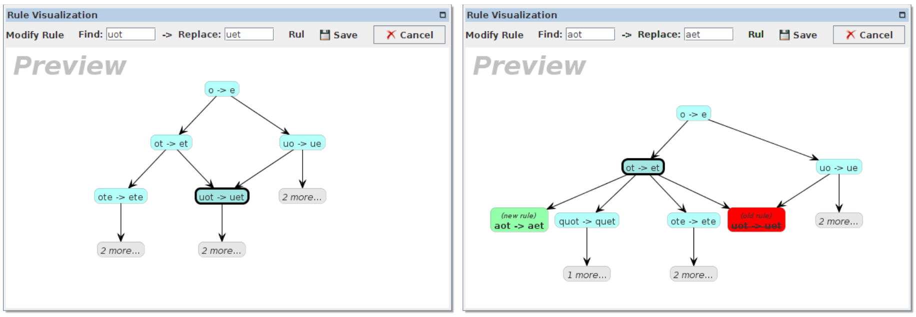
1. The component *Evidences* (top left) enables building and editing the training set that consists of pairs of contemporary spelling and corresponding spelling variation. An automatic evidence construction [Ernst-Gerlach/Fuhr 2010a/b] has been integrated. As result the user receives a list (see figure 1) where he can accept evidences. Experimental results have demonstrated the feasibility of this approach. Thus a training set can be constructed with much less initial effort. Afterwards the user can build evidences based on the remaining unknown words. The system supports this step by offering a list with suggestions from the spell-checker.

2. The component *Historic-Text* (top right) offers context information for spelling variants in an abridgement of the text in order to clarify the meaning of the word.

The *Rule-Modification* is also subdivided into two components:

1. The *Rule-Selector* (bottom left) gives an overview over the rule set and enables a search for certain rules by means of filters.

2. The component *Rule-Visualization* (bottom right) enables the user to look and work at certain rules as well as enter new rules. In order to facilitate the modification of rules a preview-mode demonstrates the variations before they are really applied (see figure 3).



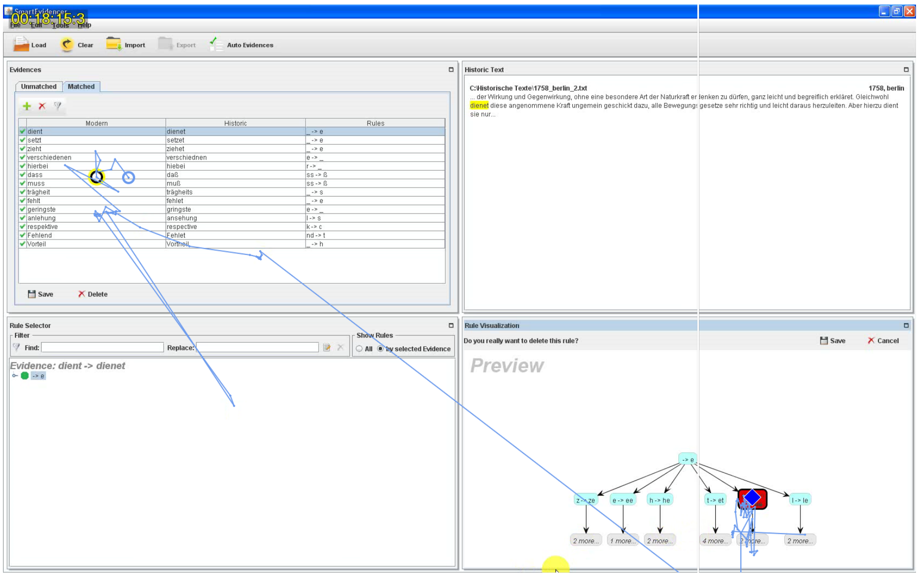
Modify Rule Find: uot -> Replace: uet

Modify Rule Find: aot -> Replace: aet

Figure 3: Preview-Mode

The usability of the *RuleGenerator* application has been checked within an eyetracker-supported evaluation. The application was looked upon favourably in general. The users had no essential problems when they worked on the given tasks. There only have been minor comments e. g. about to less explicit icons. The eyetracker data (see Figure 4) have been helpful for analysing the user behaviour. According to the problems the *RuleGenerator* has been enhanced.

The presented *RuleGenerator* which automatically detects evidences and generates rules provides users with much more comfort to access historic documents. Thus the tool enables retrieval in texts with non standard spelling with a very flexible approach since the user can influence the process according to his needs regarding recall and precision.



	Modern	Historic	Rules
✓	dient	dienet	_ -> e
✓	setzt	setzet	_ -> e
✓	zieht	ziehet	_ -> e
✓	verschiedenen	verschiednen	e -> _
✓	hierbei	hiebei	r -> _
✓	dass	daß	ss -> ß
✓	muss	muß	ss -> ß
✓	trägheit	trägheits	_ -> s
✓	fehlt	fehlet	_ -> e
✓	geringste	gringste	e -> _
✓	anlehung	ansehung	l -> s
✓	respektive	respective	k -> c
✓	Fehlend	Fehlet	nd -> t
✓	Vorteil	Vortheil	_ -> h

Evidence: dient -> dienet

C:\Historische Texte\1758_berlin_2.txt 1758, berlin
... der Wirkung und Gegenwirkung, ohne eine besondere Art der Naturkraft erdenken zu dürfen, ganz leicht und begreiflich erkläret. Gleichwohl dienet diese angenommene Kraft ungemein geschickt dazu, alle Bewegungsgesetze sehr richtig und leicht daraus herzuleiten. Aber hierzu dient sie nur...

Do you really want to delete this rule?

Figure 4: Scanpath of a proband

References

Awakian, A. (2010). Development of a user-interface for an interactive rule development. Master thesis, University of Duisburg-Essen

Baron, A.; Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. Proc. of the „Postgraduate Conference in Corpus Linguistics“. Aston University, Birmingham

Cendrowska, J. (1987). An Algorithm for Inducing Modular Rules. „International Journal on Man-Machine Studies“. Volume 27, Number 4, pp. 349-370

Ernst-Gerlach, A.; Fuhr, N. (2006). Generating Search Term Variants for Text Collections with Historic Spellings. Proc. of the "Advances in Information Retrieval - 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006"

Ernst-Gerlach, A.; Fuhr, N. (2010a). Advanced Training Set Construction for Retrieval in Historic Documents. Proc. of the "Sixth Asia Information Retrieval Societies Conf. (AIRS 2010)"

Ernst-Gerlach, A.; Fuhr, N. (2010b). Semiautomatische Konstruktion von Trainingsdaten für historische Dokumente. Proc. of the "Information Retrieval 2010 Workshop“ LWA, Kassel, Germany

Hauser, A.; Heller, M.; Leiss, E.; Schulz, K. U.; Wanzeck, C. (2007). Information Access to Historical Documents from the Early New High German Period. Proc. of the „International Joint Conference on Artificial Intelligence (IJCAI-2007) Workshop on Analytics for Noisy Unstructured Text Data“ Hyderabad, India

Korbar, D. (2010). Visualisierung von Regelstrukturen und -Modifikationsmöglichkeiten für die Suche in Texten mit nicht-standardisierter Rechtschreibung. Dipomarbeit, University of Duisburg-Essen

Pilz, T. (2009). Nichtstandardisierte Rechtschreibung -Variations-modellierung und rechnergestützte Variationsverarbeitung. Dissertation, University of Duisburg-Essen