Martin-D. Gleßgen (Zürich)
Matthias Kopp (Tübingen)

# Linguistic annotation of texts in non-standardized languages: the program procedures of the tool PHOENIX

Cet aperçu met en relief les principes de fonctionnement du lemmatiseur qui constitue le noyau du programme PHOENIX. Celui-ci a été développé pour l'annotation semi-automatique de textes écrits en ancien français, mais il peut être appliqué à toute autre forme de langue médiévale ou non-standardisée en général. Sa particularité réside notamment dans la prise en compte des exigences philologiques contraignantes qu'implique le traitement des textes médiévaux.

## 1. Aims of the tool and intentions of the present overview

The aim of the PHOENIX Project is to provide an analytical tool for historical and comparative linguistics. PHOENIX is based on parameterized modules of TUSTEP (*TUebinger System von TextverarbeitungsProgrammen*; see also Matthey, in this volume) and on procedures written in the TUSTEP scripting language. This paper intends to define on the one hand the philological and linguistic needs met by the tool. On the other hand information on the XML-structures used and maintained behind the scene shall be given. Furthermore the means by which these data are managed shall be demonstrated. We will thus try to show how text data represented in an XML-structure are enriched and augmented with the help of TUSTEP procedures and several further corresponding XML-structures.

## 2. Philological and linguistic requirements

The PHOENIX project is concerned with the management of texts written in non-standardized language varieties available in the corpora of medieval European *manuscript culture*. This means that special attention has to be paid to the handling of orthographical, morphological and lexical variances. It means also that the textual data have a particularly complex structure due to

the fact that the texts dealt with and also their transmission require emendations and comments: their transmission through different copies implicates divergences which have to be respected and mentioned by the editor. PHOENIX is therefore designed for highly segmented data with, for example, critical apparatus and great varieties in hierarchical structure.

A second characteristic of PHOENIX is the modeling of the consecutive steps required for a philological and linguistic analysis of ancient texts. The tool allows lexicological, graphematical and morphological annotation of untagged textual data. This may be done either without any comparative terms or by introducing a thesaurus of inflected or uninflected forms (cf. 4. *infra*). PHOENIX allows thus to manage an interactive relationship between a textual database and a corresponding interpretative database (being graphical, onomastic, graphematic, morphological or syntactic). It enables queries based on such annotated and enriched data, in order to identify linguistic changes in terms of time, space, sociological groups and textual genres. The emulation of a traditional *modus operandi* guarantees for the quality of the analysis regarding diachronic linguistics.

Eventually PHOENIX permits export of textual data in either print or digital form, i.e. an "electronic edition" (cf. fig. 1 and Gleßgen 2003). On the other hand the acquired interpretative information will be exported as glossary or dictionary.

Parchemin jadis scellé sur simple queue; 58x141
AD MM H 1244, fonds de l'abbaye de Salival

Ecr.: semi-onciale archaïque, frustre, statique, très lisible; *s-* long systématique
Langue: latinisme *chapitle* (2)

**1** Conue chose soit a-toz **2** q*ue* li abes *et* li chapitles de Salinvas · at laissié a Wirion / *et* Huillon, les dous freres de Gev*er*lise, les anfanz Bertran Bacheler, **3** ·XIII· jor/nas de t*er*re treisse · en la fin de Gev*er*lise · *et* a lor oirs · **4** p*ar*mi ·XIII· d*eniers* de cens · *et* / ·II· himas de blef · l'un d'avoine · l'autre de froment · **5** *et* s'il ne paievent a jor // nomei a la feste sent Remi · a Giv*er*lise, en la maison de Salinvas[1] · q*ue* l'on se tan/roit a la terre · *et* ce q*ue* sus averoit · **6** Si est ensi devisee · q'au Tramble en / at ·IIII· jornas · un p*ar* lui[2] · *et* ·III· ensemble · **7** *et* en la voie de Hignicort en at / V· jornas, ·II· d'une part *et* ·III· d'autre · **8** *et* en la voie de Marsal ·II· jornas · / aprés la t*er*re les Vowes[3] · **9** *et* en la voie de Donnereis · as Genoivres · en at // ·II· jornas · **10** Ci at mis li abes *et* li covenz de Salinvas son sael · en tesmoig/nage de verité · **11** l'an q*ue* li miliaires corroit p*ar* ·M· *et* CC· *et* XXXIIII· anz ·

[1] L'abbaye possède donc une maison à Juvelize. [2] Probablement *Wirion*, le premier frère nommé dans le texte. [3] *Les Vowes* ou des *Vowes*?

Fig. 1. Printout of a charter (detail): metainformation and text of the Old French charter; scholarly comments referenced by footnote-numbers are given "apparatus-like" below.

## 3. Structure of source data

PHOENIX has been developed and was used until nowadays for the treatment of Old French charters, i.e. text data written in a non-standardized language with high graphematic, morphological and lexematic variance. These data raise numerous questions in the field of historical linguistics, concerning, for example, the elaboration of written French language, the identification of scriptoria, the characteristics of discourse in various specific vocabularies (law, agriculture, seigneurial relationships) or the grammatical uses proper to specific textual genres (cf. Gleßgen 2005).

To answer those questions by means of IT it is necessary first of all to store the textual data (and the metadata available at this time) in an appropriate structure. For processing with PHOENIX the source text has to be available as valid XML, i.e. conformant with a PHOENIX-DTD and a corresponding schema (cf. fig. 2): meta-information concerning time, date, content, author are stored in elements at the top (= element AN); the text of the charter is contained in the element TXT; substructures representing classification of components or scholarly comments as well as structuring elements introduced by the medieval scribe or abbreviations are marked up by subordinate elements (e.g. DIV N="x" for a semantic entity, FUL for a footnote, ABR for an abbreviation).

```
<f>Parchemin jadis scellé sur simple queue; 58x141</f>
<l>AD MM H 1244, fonds de l'abbaye de Salival</l>
<ec>semi-onciale archaïque, frustre, statique, très lisible;
<abr>s-</abr> long systématique </ec>
<met>latinisme <abr>chapitle</abr> (2)</met></an>
<txt>
<pub>
  <div n="1"><maj>C</maj>onue chose soit a-toz</div></pub>
<exp>
  <div n="2"> q<abr>ue</abr> li abes <abr>et</abr> li chapitles
    de Salinvas /. at laissié a Wirion
    <zw/> <abr>et</abr> Huillon, les dous freres de
    Gev<abr>er</abr>lise, les anfanz Bertran Bacheler,
  </div>
  <div n="3">/.XIII/. jor<zwt/>nas de t<abr>er</abr>re
    treisse,//. en la fin de Gev<abr>er</abr>lise /.
    <abr>et</abr> a lor oirs,//.
  </div>
  <div n="4"> p<abr>ar</abr>mi /.XIII/. d<abr>eniers</abr> de
    cens /. <abr>et</abr>
    <zw/> /.II/. himas de blef,//. l'un d'avoine,//. l'autre de
    froment;//.
  </div>
  <div n="5"> <abr>et</abr> s'il ne paievent a jor
    <zw/> nomei a la feste sent Remi,//. a
    Giv<abr>er</abr>lise, en la maison de Salinvas
    <ful>L'abbaye possède donc une maison à Juvelize.</ful>,//.
     q<abr>u</abr>e l'on se tan<zwt/>roit a la terre /.
     <abr>et</abr> ce q<abr>ue</abr> sus averoit.//.
</div>
```

Fig. 2. XML-structured source data.

## 4. Annotation procedure

The first step of the analysis is to identify and to collect all forms having a similar function in the text corpus. This is achieved with the help of a list of all forms – more precisely all graphical words – generated by means of the powerful built-in features of TUSTEP. This means that user-defined equivalents may be taken into account when this list is generated. It is therefore possible to subsume varying forms in the resulting entries. Rules describing linguistic phenomena like reduplication (*abbé* = *abé*), graphematical redundancy (*abbaye* = *abbaie*), sound shift (*abbé* = *abbei* = *abbeit*) or inflectional variance (*venons* = *venez*) enable automatic collection of forms in one entry; to express it in terms of data structure: to collect them in one element (cf. fig. 3: two fragments showing different strings subsumed under the lemma "abbé"; the elements <ah> and <rh> contain absolute and relative frequencies; each occurrence is connected with a numeric address describing its position in the source data).

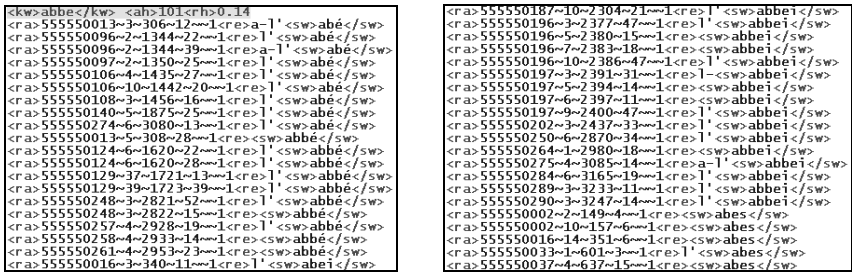


Fig. 3. List of word forms

The definition of equivalents respects the diasystematic qualities of the text data (era, region, textual genre) and these definitions may ultimately be applied to the respective parts of a complete corpus. This approach simultaneously allows the development as well as the verification of hypotheses concerning the characteristics of graphematical or morphological variance.

In the next step a qualification of the obtained collection is accomplished. This is necessary since the words (or occurrences) put together by means of either graphical identity or user-defined equivalence could differ from a linguistic point of view: they may belong to different word classes or lexems. On the other hand different entries may also belong to one single lexem.

The definition of groups and the assignment of one occurrence to a group has to be done interactively by a scholar with the knowledge of the treated language variety. A graphical interface enables to select and to pool particular forms describing for example grammatical aspects or grapho-phonetic char-

acteristics.[1] Decisions may be achieved with respect to the permanently available context where an occurrence derives from. Described in terms of data structures and their relation this step implies:

- in the first place a markup of each processed occurrence providing a base for all future reference (= WN N="x", cf. fig. 4);
- in the second place the definition of a further component connecting the meta-information, that is, the group, defined now with one or more particular occurrences (e.g. LEX, cf. fig. 5).

```
<f>Parchemin jadis scellé sur simple queue; 58x141</f>
<l>AD MM H 1244, fonds de l'abbaye de Salival</l>
<ec>semi-onciale archaïque, frustre, statique, très lisible;
<abr>s-</abr> long systématique </ec>
<met>latinisme <abr>chapitle</abr> (2)</met></an>
<txt>
<pub>
  <div n="1"><maj>C</maj>onue chose soit a-toz</div></pub>
<exp>
  <div n="2"> q<abr>ue</abr> li <wn n="1">abes</wn>
    <abr>et</abr> li chapitles
    de Salinvas /. at laissié a Wirion
    <zw/> <abr>et</abr> Huillon, les dous freres de
    Gev<abr>er</abr>lise, les anfanz Bertran Bacheler,
  </div>
  <div n="3">/.XIII/. jor<zwt/>nas de t<abr>er</abr>re
    treisse,//. en la fin de Gev<abr>er</abr>lise /.
    <abr>et</abr> a lor oirs,//.
  </div>
  <div n="4"> p<abr>ar</abr>mi /.XIII/. d<abr>eniers</abr> de
    cens /. <abr>et</abr>
    <zw/> /.II/. himas de blef,//. l'un d'avoine,//. l'autre de
    froment;//.
  </div>
  <div n="5"> <abr>et</abr> s'il ne paievent a jor
    <zw/> nomei a la feste sent Remi,//. a
    Giv<abr>er</abr>lise, en la maison de Salinvas
    <ful>L'abbaye possède donc une maison à Juvelize.</ful>,//.
     q<abr>u</abr>e l'on se tan<zwt/>roit a la terre /.
      <abr>et</abr> ce q<abr>ue</abr> sus averoit.//.
</div>
```

Fig. 4. Occurrence marked up in XML-structured source data: the string “abes” has been processed; it is marked up as WN-element whose attribute *n* refers to an entry in the index file

```
<wn>1</wn>
<src>abes</src>
<lex f="">abbe</lex>
<nam f=""/>
<graph/>
<morph f=""/>
<sem f=""/>
<synt/>
<varia/>
```

Fig. 5. Entry in an index file

Each form processed in this step receives a further tagging as a WN-element (*wn* meaning ‘word number’) with an attribute *n* whose value is an ID. The enrichment of the source text is done by connecting selected components with an entry in an index file (cf. fig. 5 above). At any time the index file may be

1 The morphological and lexematic annotations can also be based on comparative resources. As soon as a thesaurus of the analyzed language variety exists, one could identify word classes or lemmata thanks to a complementary tool (cf. Stein / Gleßgen, in this volume); PHOENIX is then able to disambiguate equivocal / dubious results (cf. *ib.*).

integrated into the source data. In the course of work on a corpus the index files serve as reference.

When this process is finished, all relevant items from the source data are identified and qualified; word forms estimated identical (while varying for example in spelling) are grouped together. The supplementary information is stored in an index file connected with the source file.

## 5. Generation of a lexicographic description

In the next step the information added up to this point gets refined: the groups of occurrences are described more specifically. The entries of the index file are enriched with further morphological and semantic qualities. This means that subdivisions have to be established. Each of them may, for example, contain forms being identical in spelling but different in meaning (polysemy) or may cover forms with different grammatical functions in the text (nominal and verbal inflection); furthermore, in the process of refining, an occurrence may move from one subdivision to another or to another group.

The description of these divisions and subdivisions and the respective occurrences are stored in a third XML-structure, the "lexicographic file", reflecting the structure of a traditional lexicographic entry.

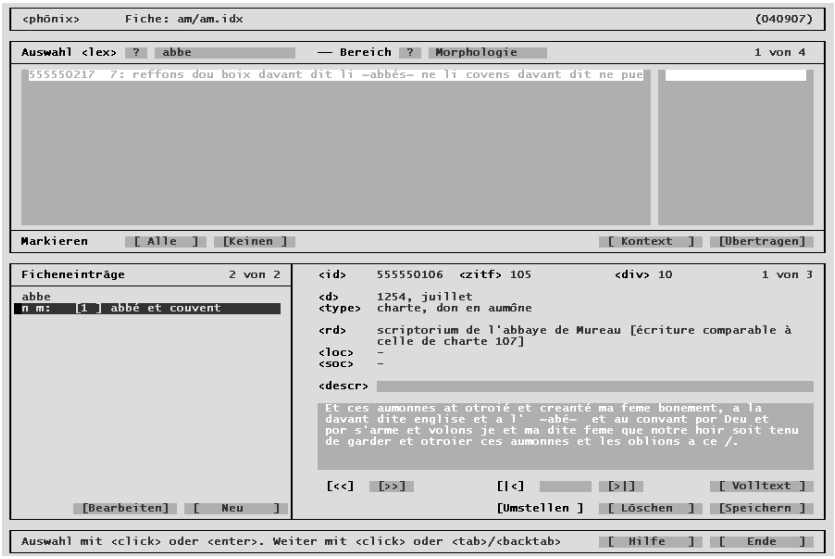


Fig. 6. Interface for the refined description of occurrences

A further interface allows for parallel access to the source file and the index file while the lexicographic information gets stored and maintained in the lexicographic file mentioned before (cf. fig. 6 on the preceding page). The aim of the parallel management of three different files (txt, idx, fiche) is to maintain an interactive link between textual and interpretative data bases.

## 6. Evaluative aspects

Research on non-standardized languages requires the integration of the analysis of a text with its edition. These two steps overlap here, while they are separated when texts of contemporary language are studied. The close relation between source data and lexicological (interpretative) data base enables modifications of the transcribed text while the analysis is in progress (e.g. word separations, accents, identification of paleographically indistinct forms as *m* = *ni* = *ju* = *iv*). This is one of the essential characteristics by which PHOENIX differs from other tools providing comparable features.

Moreover, the internal format based on XML-standards guarantees the possibility to export data into other applications. The textual data base, encoded in an XML-schema, can be transformed into a digital as well as a printed edition, which corresponds to the high standards of typographical tradition for scholarly editions.

PHOENIX provides features that are partially covered by other tools, such as simultaneously processing textual and interpretative data bases or performing lemmatization or else identifying as well as quantifying graphematical or morphological items. However, there is no tool able to manage this singular combination of features (managing complex textual data, methodological procedure of traditional *modus operandi*, interactive relationship between textual and interpretative data bases, consequent application of XML-structures) like PHOENIX does; the way PHOENIX allows to operate these is unique. Finally TUSTEP – the PHOENIX programming language – allows the processing of very large corpora.

There are other excellent tools developed for corpus linguistics, such as GATTO (for the OVI-TLIO; cf. also Bosco / Bazzanella, in this volume), STELLA (for the TLF-ATILF; cf. also Gerner, in this volume), COSMAS-I and -II (for IDS / Duden; cf. Bodmer *et al*. 2002) and the tools proposed by the SIL. But for the results required by our interests, those could not substitute PHOENIX. As a matter of fact – and independently from the fact that they are not designed for text editions –, these tools bear the following inconveniences:

- they all work on given texts that cannot be modified while operation is in progress;
- they do not necessarily guarantee an interactive relation between textual and lexicological data bases (the most evolved in this way is GATTO; STELLA and COSMAS bear this relation only up to a certain point);

- STELLA and COSMAS are nearly unavailable for scholars; STELLA can only be used for institutional collaboration; IDS partly enables the integration of texts by the users; SIL tools (less interesting for older texts editions) are available, but not free; only GATTO is free and easy to download;
- the sources of all programs are inaccessible, which does not allow changes in their functionalities; PHOENIX is conceived as an Open-Source tool for scientific purposes;
- the specific functionalities of PHOENIX are elsewhere not developed in the same systematic way; this concerns specifically the various diasystematic qualifications (the other tools integrate the textual genre, but e.g. the entity of a scriptorium is not existent) or the modular definition of graphematic or morphological equivalences;
- the elaboration of onomastical, graphematical and morphological data bases besides the lexicological one is not provided by any tool as a basic application; they all have particular strong features: COSMAS-II provides a powerful tool for the recognition of phraseologisms, STELLA makes an ingenious use of regular expressions; but none of them provides a simple functionality capable of inventorying and classifying in a parallel way the constitutive elements of language.

The two linguistic aims PHOENIX has been basically designed for, that is: the annotation and the enrichment of textual data, are now reached. The next task is to develop interfaces with other existing tools (cf. Stein / Gleßgen, in this volume). This will allow for crossed queries and quantification of linguistic changes in lexicon, grammar or grapho-phonetics. At the moment, we are studying the usefulness and the practicability of specific tools developed for XML-documents, such as X-Query which has good qualities for future standard applications.

## Bibliography

Bodmer, Frank *et al*. 2002: Von der Tonbandaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für die Arbeit mit Gesprächskorpora; in: Pusch, Claus D. / Raible, Wolfgang (eds.): *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache / Romance corpus linguistics: Corpora and spoken language*. Tübingen: Narr, 209–243.

Gleßgen, Martin-D. 2003: L'élaboration philologique et l'étude lexicologique des *Plus anciens documents linguistiques de la France* à l'aide de l'informatique; in: Duval, Frédéric (ed.): *Frédéric Godefroy. Actes du Xe colloque international sur le moyen français*. Paris: Ecole des Chartes, 371–386.

— 2005: Diskurstraditionen zwischen pragmatischen Vorgaben und sprachlichen Varietäten. Methodische Überlegungen zur historischen Korpuslinguistik; in: Schrott, Angela / Völker, Harald (eds.): *Historische Pragmatik und Historische Varietätenlinguistik*. Göttingen: Universitätsverlag Göttingen.