Editorische, lexikologische und graphematische Erschließung altfranzösischer Urkundentexte mit Hilfe von TUSTEP

Stand der Arbeiten

Von Martin-Dietrich Glessgen (Straßburg)

1. Empirischer Ansatz und Zielsetzung

Gegenstand der folgenden Betrachtungen sind die dokumentarischen Quellen der *langue d'oïl* aus Oberlothringen, die einen exemplarischen Blick auf die Entwicklung der französischen Schriftsprache zwischen dem 13. und dem 15. Jahrhundert erlauben. Die Darstellung schließt sich an den längeren Beitrag zum vorangegangenen Urkundenkolloquium an (Glessgen 2001) und soll den Fortgang meiner Arbeiten zum altfranzösischen Geschäftsschrifttum dokumentieren, ohne neuerliche Hinweise auf die bereits dargestellte Sekundärliteratur.

Als empirischer Ansatzpunkt für ein weiteres Studium der zahlreichen oberlothringischen Archivquellen sowie für jenes der ältesten altfranzösischen Urkunden insgesamt dient die Sammlung der ältesten in den *Archives Départementales de Meurthe-et-Moselle* in Nancy aufbewahrten Originalurkunden nach der soliden maschinenschriftlichen Edition von Michel Arnod (Arnod 1974). Dieses Korpus von 290 Urkunden (1231–1265) wird nun zur Publikation in der Reihe *Les plus anciens documents linguistiques de la France* vorbereitet, was einzelne Korrekturen erforderlich macht; zudem werden engere Editionskriterien angelegt und ein ausgeweitetes Regest sowie ein Glossar erstellt.

Die hier erarbeiteten Modelle sollen unmittelbar Verwendung für die elektronische Edition der Gesamtreihe der *Plus anciens documents* finden. Nach dem zu frühen Tod von Jacques Monfrin haben Françoise Vielliard und Olivier Guyotjeannin sich dieser Reihe angenommen und sich der Vorstellung einer elektronischen Edition und Auswertung angeschlossen, die wir nun gemeinsam verfolgen.

Die empirische und interpretative Durchdringung des altfranzösischen Geschäftsschrifttums verfolgt ein zweites, komplementäres Fernziel: Die Erarbeitung eines informatischen Editions- und Analyseinstruments, mit dem ältere romanische – oder auch anderssprachige – Texte bearbeitet werden

können. Die handelsüblichen Softwareprodukte sind hierfür nicht ausreichend und können es angesichts ihrer kommerziellen Orientierung auch nicht sein. Es soll ein verallgemeinerbares Modell entstehen, das die traditionelle Philologie und Lexikologie mit den Möglichkeiten der informationstechnologischen Textaufbereitung und Sprachuntersuchung verknüpft.

In den vergangenen zwei Jahren habe ich in Zusammenarbeit mit Matthias Kopp (Tübingen) eine feste, aber zugleich offene Struktur der Textdaten entwickelt, die unterschiedliche editorische Darstellungen erlaubt (2./3.). Auf dieser Grundlage entstand dann mit Hilfe des Programmpakets TUSTEP eine Abfrageroutine, die sowohl die lexikalische (4.) als auch die graphematische Auswertung (5.) vorbereitet. Der folgende Überblick konzentriert sich als Werkstattbericht auf die technischen Aspekte der Auswertung, da die systematische sprachwissenschaftliche Arbeit erst auf der Basis eines entwickelten Programms zum Tragen kommen kann.

2. Edition

Die schon kurz dargelegten Editionsprinzipien (GLESSGEN 2001, 276) bemühen sich um einen Kompromiss zwischen 'alter' und 'neuer' Philologie, also zwischen der Aufgabe des Herausgebers, den Text dem modernen Leser nahe zu bringen und verständlich zu machen, und den Anforderungen der Sprachwissenschaft, die eine möglichst ungebrochene maschinenlesbare Reproduktion des Originals benötigt. Die folgenden kombinierten Editionskriterien sollen, unter Nutzung der Möglichkeiten elektronischer Medien, beiden Zielvorgaben Rechnung tragen:

1. Zeichensetzung: Die Originalzeichensetzung wird beibehalten; sie erscheint wie in den Originalen auf halber Höhe der Linie. Zusätzlich wird eine moderne Zeichensetzung auf der Grundlinie eingeführt, die dem Leser das Verständnis erleichtern soll (s. u. Anm. 6).
2. Großschreibung am Wortanfang: Die Original-Großschreibung wird – soweit sie erkennbar ist[1] – durch Fettdruck dargestellt. Parallel dazu werden Großbuchstaben nach modernen Regeln eingeführt.

[1] Die Umsetzung der in Größe und Form stark varianten Anfangsbuchstaben (cf. BROMM 1995) in die binäre Opposition, die sich erst später als Prinzip verfestigt hat, bedeutet in jedem Fall einen starken Eingriff in die Realität der mittelalterlichen Formen. Die Handschriften weisen – gleichzeitig oder alternativ – Markierungen durch Größe, Duktus und Buchstabenschmuck auf, die in ihrer Ausgeprägtheit und Kombination sowie im Vergleich mit den übrigen Formen desselben Texts zu gewichten sind. Ein dazu entwickeltes Raster, das auch die Position im Text (erste oder

3. Textgliederung: Der Zeilenwechsel im Original wird angegeben, und die Originalzeilen werden durchgezählt (im Fünferschritt am Rand). Hinzu tritt eine durchnumerierte inhaltlich-syntaktische Gliederung, die eine inhaltliche Interpretation darstellt und dem besseren Verständnis dient.
4. Schriftformen: Die Transkription verzichtet auf die Wiedergabe paläographischer Details wie Buchstabenformen, präzise Formen der Zeichensetzung und der Diakritika, Striche oder Tilden am Zeilenrand. Dafür erscheint eine photographische Abbildung des Originals (mit Vorder- und Rückseite): Weitergehende Analysen der Text- und Schriftgestalt sind also möglich.[2]
5. Kontextuelle Verankerung: Jede Urkunde wird von einem umfangreichen Analyseteil begleitet mit einem Regest, Angaben zu Textsorte, Inhalt, auftretenden Personen, mutmaßlichem Aussteller sowie dem Lager und der Texttradition. Entscheidend für die diasystematische Verankerung der Urkunden ist dabei, neben der Datierung, der Aussteller.

Das am Beispiel der folgenden zweitältesten Urkunde des Korpus dargestellte Mischsystem zeigt eine mögliche Verwirklichung der Editionskriterien:

---

letzte Zeile, Anfang der zweiten Zeile, nach einem Satzzeichen) und die Wortsemantik (Titel, Institution) berücksichtigt, ergibt bisher tragfähige Ergebnisse.

[2] Dieser Punkt wird zur Zeit kontrovers diskutiert. Nicht nur die Veranstalter des Kolloquiums fordern noch engere Transkriptionskriterien. Gewiß ist – wie Holtus / Völker 1999 darstellen – die Opposition zwischen langem und rundem *-s-* für die nordfranzösischen Urkunden der zweiten Hälfte des 13. Jahrhunderts ein Leitmerkmal, das Lokalisierungshinweise geben kann. Doch erbringt die Auswertung graphematischer Eigenarten denselben Ertrag (vgl. infra Abschnitt 5.), der dann unmittelbar von sprachwissenschaftlicher Relevanz ist. Zudem können graphematische Elemente beim augenblicklichen Kenntnisstand im gesamten Netzwerk der graphophonetischen Relationen des Altfranzösischen beurteilt und gedeutet werden. Das ist heute, zumal bei zeit- und raumübergreifenden Studien für Buchstabenformen noch nicht möglich. Jede diesbezügliche Arbeit (etwa die genannte Studie von Bromm oder auch, für die Zeichensetzung, Llamas Pombo 1999 mit weiterführender Bibliographie) zeigt, welche Aufgaben hier noch auf die Forschung warten. Es ist nicht verwunderlich, aber bezeichnend, daß sich die paläographisch orientierten Arbeiten nur ganz am Rande mit den hier im Vordergrund stehenden sprachinternen Fragen beschäftigt: Forschung kann und darf komplementär sein.

002

1234 (25 mars–31 décembre) ou 1235 (1er janvier– 24 mars)

Type de document: charte: acensement de terres
Objet: *L'abbé et le chapitre de Salival acensent à Wirrion et Houillon treize journaux de terre au finage de Juvelize contre un cens de treize deniers et deux hémines de grain.*

Auteur: non annoncé
Disposant: abbaye de Salival
Sceau: disposant
Bénéficiaire: disposant
Autres acteurs: Wirrion et Houillon, paysans de Juvelize

Rédacteur: scriptorium de l'abbaye de Salival

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*?

Die neufranzösische Übersetzung der Urkunde ist nach den mittelalterlichen Strukturmerkmalen – also Zeichensetzung und Großbuchstaben – gegliedert, was deren Pertinenz für inhaltliche Belange neuerlich zeigt (vgl. Glessgen 2001, 279f.).

[majuscule = début]	Qu'il soit connu à tous que l'abbé et le chapitre de Salival
[point = résumé de l'acte juridique]	ont cedé à Wirion et à Huillon, deux frères demeurant à Juvelize et fils de Bertran Bachelier, treize journaux de terre en friche (= *treisse*)[3]
[lieu]	dans le territoire de Juvelize;
[durée]	la cessation concerne aussi leurs héritiers;
[cens]	ceci pour la valeur de treize deniers de cens (annuel)
[cens en nature]	et deux hémines de grains, l'une d'avoine, l'autre de froment.
[condition]	Au cas où ils ne payeraient pas le jour nommé à la fête de Saint Rémy
[lieu de paiement]	dans la maison (du chapitre) à Salival
[conséquence]	l'on prendrait comme gage la terre
[spécification]	et ses fruits.
[point + majuscule: description du terrain]	(Le terrain) est réparti de la manière suivante:
[première parcelle]	au lieu-dit Tramble se trouvent quatre journaux,
[répartition entre les frères]	dont un pour (Wirion) seul,
[id.]	les trois autres (pour les deux frères) en indivis;
[deuxième parcelle]	et près du chemin qui mène à Hignycourt se trouvent cinq[4] journaux, deux pour un frère, trois pour l'autre;
[troisième parcelle]	et près du chemin qui mène à Marsal deux journaux,
[localisation précise]	derrière le terrain dit «des Vowes»;[5]

[3] Das Wort (<anfrk. **threosk* 'Brachland') ist nordostfranzösisch (wallon., flandr., champ., lothr.); es erscheint in den Varianten *trie(z)* 1257, *trieu, trihe, triexhe, triot* sowie – nur in Lothr. – *treixe* 1340 (cf. FEW 17,400b mit der unglücklichen Definition 'terrain inculte' [statt 'terre en friche']). Unsere Urkunde liefert den Erstbeleg für den Regionalismus. Darüberhinaus leitet die (im FEW nicht vermerkte) Verbindung *terre treisse* auf die adjektivische Verwendung hin, die charakteristisch für Lothringen ist und bisher nur in modernen Dialekten nachzuweisen war (cf. FEW «adj. Meuse *trîce,* Moselle *triχ,* Metz *trɛχ»).*

[4] Der Punkt vor der Ziffer fehlt aufgrund des an dieser Stelle erfolgten Zeilenwechsels.

[5] Oder: le terrain appartenant à la famille Vowes.

[quatrième parcelle]	et près du chemin qui mène à Donneroy
[localisation précise]	au lieu-dit Genoivres
[suite]	se trouvent deux journaux.
[point + majuscule: *corroboratio*]	L'abbé et le couvent de Salival ont mis leur sceau,
[formule de *corroboratio*]	en témoignage de vérité
[*datatio*]	en l'année 1234.

Die 'neue Philologie' entfaltet nun ihre eigentliche Wirksamkeit in der elektronischen Edition, bei der jeder Benutzer die ihn interessierende Darstellungsform wählen kann; also:
- eine Darstellung nach dem obigen Mischsystem;
- eine Darstellung nach den stärker interpretativen Prinzipien der DocLing=Fr, mit der alleinigen modernen Zeichensetzung und Großschreibung;[6]
- eine eher diplomatische Darstellung, mit alleiniger Zeichensetzung und Großbuchstaben des Originals, geordnet nach den Originalzeilen zum unmittelbaren Vergleich mit der Photographie; dabei kann die Einführung von Apostrophen und Worttrennung rückgängig gemacht werden, was die Agglutination der Determinanten optisch hervorhebt.

Aufgrund der guten Möglichkeiten, ein TUSTEP-File mittels einer CGI-Schnittstelle in eine browserfähige Darstellung zu überführen, ruht diese elektronische Edition auf einer sicheren Grundlage.[7]

3. Kodierung der Textdaten

Alle genannten Sichten auf das Dokument werden unmittelbar und automatisch aus einem einzigen Grundtext abgeleitet. Dieser ist auf der Grundlage einer XML-Struktur systematisch kodiert, was sowohl die nötige Langzeitsicherung garantiert als auch die gewünschten sprachwissenschaftlichen und inhaltlichen Abfragen ermöglicht. In dieser Kodierung hat die Urkunde von 1235 folgende Form:

```
<gl>
<t type =”123“/>
```

[6] Bei der Kodierung des Textes müssen zu diesem Zweck neben den mittelalterlichen Interpunktionszeichen konsequent alle modernen Interpunktionen angegeben werden, also gegebenenfalls auch an derselben Stelle.

[7] Eine exemplarische Browserdarstellung wurde beim Trierer Vortrag präsentiert, jedoch noch ohne die Programmierung der Wahlschalter.

```
<id>555550002</id>
<zitf>002</zitf>
    <an>
    <nom>002</nom>
    <d>1234 (25 mars–31 décembre) ou 1235 (1er janvier–24 mars)</d>
    <d0></d0>
    <type>charte: acensement de terres</type>
    <r>L'abbé et le chapitre de Salival acensent à Wirrion et Houillon treize
    journaux de terre au finage de Juvelize contre un cens de treize deniers et
    deux hémines de grain.</r>
    <aut>non annoncé</aut>
    <disp>abbaye de Salival</disp>
    <s>disposant</s>
    <b>disposant</b>
    <act>Wirrion et Houillon, paysans de Juvelize</act>
    <rd>scriptorium de l'abbaye de Salival [rdp: néant; com: soustraitance?]
    </rd>
    <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 Salin-
        vas /. 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 Juveli-
        ze.</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>
<par>
<div n=6> <maj>S</maj>i est ensi devisee /. q'au Tramble en
<zw/> at /.IIII/. jornas /. un p<abr>ar</abr> lui <ful>Probablement
<abr>Wirion</abr>, le premier frère nommé dans le texte.</ful> /.
<abr>et</abr> /.III/. ensemble /.</div>
<div n=7> <abr>et</abr> en la voie de Hignicort en at
<zw/> V/. jornas, /.II/. d'une part <abr>et</abr> /.III/. d'autre /.
</div>
<div n=8> <abr>et</abr> en la voie de Marsal /.II/. jornas /.
<zw/> aprés la t<abr>er</abr>re les Vowes <ful><abr>Les Vowes
</abr> ou des <abr>Vowes</abr>?</ful> /.</div>
<div n=9> <abr>et</abr> en la voie de Donnereis /. as Genoivres /. en
at
<zw/> /.II/. jornas /.</div></par></exp>
        <cor>
<par>
<div n=10> <maj>C</maj>i at mis li abes <abr>et</abr> li covenz de
Salinvas son sael /. en tesmoig<zwt/>nage de verité /.</div></par>
</cor>
        <dat>
<div n=11> l'an q<abr>ue</abr> li miliaires corroit p<abr>ar</abr> /.
M/. <abr>et</abr> CC/. <abr>et</abr> XXXIIII/. anz /.</div></dat>
</txt>
</gl>
```

Die verwendeten – in spitzen Klammern enthaltenen – Tags oder Steuerzeichen zerfallen in drei Gruppen:

– zunächst die allgemeinen Strukturelemente: <gl> ~ </gl> zur Rahmung einer Entität, der noch zu erläuternde Texttyp (<type = …>), ein Numerus currens (<id>) und eine Zitierform (<zitf>), die insbesondere bei größeren Textmengen wichtig wird:

```
<gl>                = début d'une entité
<t type ="123"/>    = genre textuel (= détermine la structure des balises)
<id> </id>          = numéro d'identification
<zitf> </zitf>      = sigle („zitierform")
      (…)
</gl>               = fin de l'entité
```

– der Analyseteil (zwischen <an> und </an>) mit den bereits erwähnten Angaben (Überschrift = <nom>, Datum, Textgattung, Regest etc.):

<an>	= début du tableau analytique
<nom> </nom>	= nom du texte (= ici: numéro de la charte)
<d> </d>	= date
<type> </type>	= type de document
<r> </r>	= regeste
<aut> </aut>	= auteur
<disp> </disp>	= disposant (p. ex. en cas de vidimus)
<s> </s>	= sceau
	= bénéficiaire
<act> </act>	= autres protagonistes
<rd> </rd>	= rédacteur
<sc> </sc>	= scribe
<f> </f>	= forme et description matérielle de la charte
<l> </l>	= lieu de conservation
<ed> </ed>	= édition éventuelle
<ana> </ana>	= regeste ou analyse éventuelles
<ec> </ec>	= observations sur l'écriture
<met> </met>	= observations sur la langue
</an>	= fin du tableau analytique

– schließlich der eigentliche Textteil (zwischen <txt> und </txt>), mit den Kodierungen für die inhaltlichen Abschnitte ( ) und die Zeilenwechsel des Originals (<zw/>[8] bzw. <zwt/>, wenn der Zeilenwechsel zugleich eine Worttrennung impliziert). Hier sind auch die Abkürzungen und Großbuchstaben des Originals erfaßt (<abr>, <maj>, die zwei Sonderserien der Zeichensetzung sowie zwei Fußnotenserien (<fue>, <ful>):

<txt>	= début du texte
	= chiffres gras pour la structuration sémantique du texte
<zw/> ~ <zwt/>	= changement de ligne dans l'original („zeilenwechsel + trennung")
<par>	= nouveau paragraphe
/. /, /; /.. /…	= signes de ponctuation originaux
./ ,/ ;/ :/	= ponctuation exclusivement moderne
<abr> </abr>	= abréviation dans l'original

[8] Die Darstellung dieses zugleich schließenden und öffnenden Tags ist im XML-Muster unüblich, aber möglich.

<maj> </maj>	= majuscule dans l'original
<fue> </fue>	= notes éditoriales (série indiquée par a, b, c)
<ful> </ful>	= notes concernant le contenu de la charte (série indiquée par 1, 2, 3)
<v> </v>	= texte sur le verso
</txt>	= fin du texte

In der Transkription werden außerdem vom Herausgeber eingeführte Worttrennungen (Typ *a-savoir*) und Apostrophen (*d'une part*) unterschieden. Die Urkundenteile markieren weitere Tags (Intitulatio <int>, Publicatio <pub>, Narratio <nar>, Expositio <exp>, Corroboratio <cor>, Datatio <dat>).

Eine besondere Bedeutung hat der Texttyp, da er die Integration ganz unterschiedlicher Texte in ein einziges größeres Raster ermöglicht: Urkunden, Rechnungsbücher, Versromane, Prosatexte, wissenschaftliche Traktate oder religiöse Literatur. Für jeden Texttyp entsteht eine eigene Serie von Tags, die nur in bestimmten Kernelementen übereinstimmen müssen. Die numerische Angabe am Anfang (type = …), erlaubt es dem Programm, die verschiedenen Grundmuster für die Druck- oder Browserdarstellung und für die sprachwissenschaftliche Abfrage parallel zu verwalten. Dieses Konzept ist flexibler als die Kodierung nach dem Muster der TEI (*Text Encoding Initiative*), bei der die Tags von vorneherein feststehen.

Anhand dieser Grundmuster können in einem späteren Stadium verschiedenste maschinenlesbare Texte an die Abfrageroutine angepaßt und unmittelbar verglichen werden. Bisher wurden erst einige wenige Modelle exemplarisch adaptiert, etwa die acht von CINZIA PIGNATELLI (Poitiers) ausgewerteten Manuskripte des *Chevalier de la Charrette* (ed. KARL UITTI) oder die zur Zeit von MAGALI GRANDVAL bearbeitete mittelfranzösische Übersetzung des *De arte venandi cum avibus*.

4. Lexikologische Analyse

Die sprachwissenschaftliche Auswertung der Texte soll die verschiedenen Kernbereiche der Sprache berücksichtigen, Graphematik, Lexik und Onomastik, Morphologie und Syntax sowie Textlinguistik. Sowohl für quantifizierende als auch für inhaltlich orientierte Fragestellungen ist dabei stets der erste Schritt eine Gruppenbildung, bei der thematisch zusammengehörige Formen als solche erfaßt – also getaggt – werden, um dann weiterbearbeitet zu werden. Die Möglichkeiten der maschinenlesbaren Korpora

legen es nahe, bei der Gruppenbildung zunächst formale Elemente als Anhaltspunkte zu verwenden, bevor semantische Momente hinzutreten.

Die mit Matthias Kopp als erste entworfene lexikologische Routine kann methodisch als Ausgangspunkt sowohl für die graphematische und onomastische wie, mit Abstrichen, für die morpho-syntaktische Auswertung dienen. Hier wie da soll ein Programm eine rasche Ordnung und Primärbeurteilung von Wortformen erlauben und deren weitergehende Bearbeitung vorbereiten.

In der Lexik ist das beste sprachhistorisch verankerte Ordnungsprinzip für unsere Zwecke die etymologisch geleitete Lemmatisierung, auch aus kognitiver Sicht. Genetisch verwandte Formen weisen trotz ihrer formalen und semantischen Varianz eine kognitiv gesteuerte Nähe zueinander auf (vgl. Glessgen i. Dr. b).

Das Vorgehen ist im ersten Schritt relativ einfach und klassisch: Es wird ein KWIC-Index erstellt, der alle Formen der einzelnen Lexeme im Zeilenkontext mit der entsprechenden Stellenangabe widergibt ('Key Word in Context'). Eine erste Trennung erfolgt dabei zwischen Eigennamen und Wörtern, was in unserem Korpus automatisch durch die Scheidung nach Groß- und Kleinschreibung erfolgen kann; andernfalls müßte eine Scheidung durch Tags durchgeführt werden.[9]

Dieser KWIC-Index erfährt dann eine weitere Vorstrukturierung durch die Formulierung graphischer 'Äquivalenzen', was mit TUSTEP besonders leicht möglich ist und der Natur der Quellen Rechnung trägt. Die stark varianten älteren romanischen Texte weisen eine relativ große Zahl nahezu absoluter graphischer Äquivalenzen auf. In unserem Korpus steht ein doppeltes *bb* eigentlich nie in graphematischer Opposition mit einfachem *b*, und das gilt auch für die anderen Doppelkonsonanten, also *cc/c, dd/d, ff/f, ll/l, mm/m, rr/r* oder *ss/s;* auch funktional nicht differenzierende Vokaldoppelungen sind häufig (*aa/a, ee/e, ee*|[10] */é* oder *oo/o*). Diese Eigenart tritt auch in anderen Epochen und (romanischen) Sprachen als Merkmal einer nicht völlig beherrschten Schriftlichkeit zutage (vgl. z. B. Ernst / Wolf 2002).

Andere für das Programm definierte 'Äquivalenzen' sind komplizierter, haben sprachhistorisch teilkomplementäre Funktionen und können nur als Hinweise für ähnliche Funktionen aufgefaßt werden; z. B.:

---

[9] Nur Namen, die zu Anfang einer Divisio erscheinen, müssen schon jetzt eigens getaggt werden, da ansonsten Großbuchstaben zu Beginn einer Divisio automatisch als Lexeme eingestuft werden.

[10] = *ee* am Wortende.

– latinisierende Konsonantengruppen: *cq/q, ct/t;*
– Homophone: *en/an, y/i, z|/s, (n)gni/ngn/gn;*
– graphische Redundanz: *k/qu, x|/us;*
– regionale phonetische oder graphematische Varianz: *ei|/é, eir|/er, np/mp, w/g.*

Es geht hier wohlgemerkt nur um das Ziel, die Formen der Texte vorzustrukturieren. Die Anwendung der genannten Äquivalenzen auf unser Korpus reduziert die Grundvarianz auf weniger als die Hälfte. Dabei entstehen nur wenige unetymologische Zuordnungen, die im Nachhinein wieder aufgehoben werden müssen. Ein Beispiel: Unter *abé* 'Abt' erscheinen die fünf zuvor getrennten Formen *abé, abbé, abei, abbei* und *abbey,* die alle auch zu diesem Lemma gehören. Die Formen des Rektus Singular und des Obliquus Plural bilden dagegen eine eigene Serie *(abés, abbés, abez),* wie auch die Formen auf *-t (abeit, abbeit);* auslautendes *-s/-z* oder *-t* darf nicht per Äquivalenz unterdrückt werden, da sonst größere Verwirrungen an anderer Stelle entstünden. Die phonetisch nahestehenden Lexeme *abbaye* und *abbesse* erscheinen trotz der Äquivalenzen getrennt, jeweils in drei Einträgen: *abaie/abbaie, abaiez* und *abbeie* sowie *abbaesse, abasse/abbasse* und *abesse/abbesse/abeesse.* Bei einer Lemmatisierung müssen also für die drei genannten Lexeme insgesamt neun Einträge berücksichtigt werden, die zwanzig unterschiedliche Formen erfassen.[11]

Eine zu den Äquivalenzen komplementäre Vorstrukturierung erlaubt die Auswahl bestimmter Frequenzabschnitte: Vor allem die niederfrequenten Wörter sind für die lexikalische Semantik interessant; die hochfrequenten Wörter haben dagegen stärkere morphologische und syntaktische Implikationen. Bei der Lemmatisierung ist es hilfreich, die hochfrequenten Wörter gesondert von den übrigen zu behandeln.

Schließlich ist es auch möglich, die Formen des Korpus mit einer bereits morphologisch und/oder semantisch geordneten Formenliste zu vergleichen, also eine lexikonbasierte halbautomatische Lemmatisierung in Anwendung zu bringen. Eine Prozedur dieser Art erarbeiten Matthias Kopp und ich augenblicklich in Zusammenarbeit mit Achim Stein und Pierre Kunstmann (cf. GLESSGEN i. Dr. a/b). Doch setzt dies die Existenz eines Vergleichskorpus für die jeweilige Sprachform voraus, während die Formulierung von Äquivalenzen und die Frequenzauswahl schon korpusintern anwendbar sind.

[11] Der Gedanke, graphische Äquivalenzen systematisch für eine Vorstrukturierung heranzuziehen, entstand bei einem kleinen Arbeitstreffen mit Olivier Collet und Wagih Assam (Genf), die hierzu auch weitergehende Überlegungen angestellt haben.

Nach der möglichen Anwendung der verschiedenen vorgängigen Auswahlkriterien liefert unsere automatische Routine die Wort- bzw. Namenformen parallel auf drei Ebenen:
– die automatisch erzeugten Grundformen zur raschen Durchsicht;
– die nach den Grundformen geordneten Einzelformen im Zeilenkontext mit Stellenangaben;
– schließlich den Vollkontext jeder einzelnen Stelle nach Wahl.
Diese Darstellung ist sehr handlich und kommt der empirischen Arbeit des Lexikologen am Text in idealer Weise entgegen: Er kann rasch vorgeordnete Formen sichten und in jedem Fall ohne großen Zeitverlust deren größeren Kontext zu Rate ziehen.

Im nächsten Schritt können dann diese Formen weiterbearbeitet und per Hand bestimmten Lemmata zugewiesen werden. Das Programm taggt schließlich automatisch die ausgewählten Formen im Grundtext, indem es ihnen eine bestimmte Wortnummer zuweist. Die Weiterbearbeitung und Entwicklung eines Glossars erfolgt auf der Grundlage der dabei vergebenen Attribute.

Bei dieser Methode kann die per Hand gesteuerte, einmal durchgeführte Lemmatisierung jederzeit wieder aus dem Grundtext heraus generiert werden. Es besteht also während des gesamten Prozesses eine lebendige Verbindung zwischen Grundtext und Glossar. Das ist deswegen entscheidend, weil damit keine fertige Textgrundlage vorliegen muß, bevor die lexikologische Auswertung beginnt. Die Bearbeitung des Wortschatzes führt ja zwingend zu Eingriffen in die Edition, weil erst bei der Auswertung bestimmte Textpassagen ihre gültige Deutung erfahren können (cf. MÖHREN 1997). Bereits in der jetzt vorliegenden Form weist das Analyseprogramm damit eine Reihe von Qualitäten auf, die zur Zeit kein kommerzielles Produkt leistet.

5. Graphematische Analyse

Die graphematische Analyse beruht technisch auf denselben Prinzipien der Formenauswahl wie die lexikologische. Das Programm erlaubt die gezielte Abfrage einzelner als interessant eingestufter Grapheme oder Graphemkombinationen. Wie bei Lexikon und Namenschatz werden dann alle entsprechenden Formen aus dem Text in einen KWIC-Index geschrieben, der nach denselben Methoden weiterverarbeitet werden kann.

Inhaltlich eröffnet die graphematische Varianz eine für das Korpus entscheidende Perspektive: Sie erlaubt es – zusammen mit den externen Charak-

teristika wie Schriftbild und Aufbewahrungsort sowie inhaltlichen Elementen –, den Aussteller der Urkunden zu eruieren. Diese Frage ist deswegen zentral, weil die Urkunden zwar Autor, Siegler und Datum, oft auch Zeugen und Ausstellungsort nennen, nicht aber das Skriptorium, die Kanzlei oder den Schreiber, der sie wirklich erstellt hat. Die für uns beobachtbare Sprachgeschichte des Mittelalters ist aber unmittelbar an die einzelnen Skriptorien und Kanzleien und deren Schriftsprachtraditionen gebunden (cf. GLESSGEN 2001, 283–286). Jede sprachwissenschaftliche Auswertung sollte also zunächst diese Aussteller identifizieren.

Der Einsatz graphematischer Eigenarten nutzt hierbei die zirkuläre Logik jeder variationslinguistischen Analyse: Sprachliche Indizien ermöglichen die Identifizierung der Aussteller; ausgehend von den Ausstellern können dann deren sprachinterne Eigenarten im Wandel der Zeiten untersucht werden.

Bisher habe ich etwa ein Dutzend graphematischer (und morphologischer) Erscheinungen im Korpus der *Meurthe-et-Moselle* ausgewertet (etwa *nr-ndr, aule-avle-able, wa-ga (warde), avera-avra, (l)atre-(l)etre, lo-lou-le, aus/ceaus-eus-ceus).* Die Methodik kann das zuerst analysierte Beispiel lothringisch *-nr-* vs. zentralfranzösisch *-ndr-* illustrieren. Die per Programm abgefragten Formen, die *-nr-* oder *-ndr-* enthielten, wurden zunächst etymologisch geordnet. In den meisten Fällen ergaben sich lothringisch lautgerechte *nr*-Typen, die auf lat. -N'R- beruhten (174 tokens), nur selten (2 tokens) Abweichungen davon:

AD + MINOR-: *amenrie* 'amoindrie'
ADVENIRE: *avanront* 'aviendrait'
GENERU-: *genrre(s)* 'gendre'
MANERE: *menront, remanrie (-oent)*
*MINARE: *manront, demennrray*
SUBMONERE: *semonre, resemonroit* vs. *semondront* (Urk. 3, div. 28)
TENERE: *tenr- (-ont, -ai, -a, -it, -iens), tanr- (-ont, -ai, -a(t), -oie(nt)), tinrent*
VENERIS DIE-: *vanredi* (4x) vs. *vendredi* (251,7)
VENIRE: *venr- (-a, -ons, -ont, -iens), vanr(r)ont, wanront, vinr(r)ent, con-/co(u)venr(r)– (-a, -oie), revenr- (-a, -oit), revanr-*

In diese Reihe gehören auch die Formen von PRAEHENDERE, mit dissimilatorischem Schwund des *-r-* und nachfolgendem Fall des *-d-*:

penre (-r-ons/-ont/-it), panr(r)e (-r-ai/-oi(en)t), repanre (-r-ons) [= 101 Formen]
vs. *prendre* (88,14), *prandre* (16, 63; 65), *meprendre* (144,4) [= 5 Formen]

Eine Alternanz weist auch der Ortsname *Lanrecort* (128, 29) – *Landrecors* (128,6) auf. Dagegen besteht in einer Reihe von Formen keine Möglichkeit einer Opposition zwischen *nr* und *ndr*:

– Synkope bei DONARE: *don(n)roi(en)t, -ai/-a, denroit;*
– Dissimilation in ANIMA: *enreme*, *einrme* [cf. FEW 24,581b];
– Eigennamen *Conrat, -z, Henr(r)i(s), Hanr(r)i(s), Hanr(r)ion(s)* und *Lohenrreine.*

Umgekehrt weist keiner der auf lat. -NDR- oder -NG'R- zurückgehenden Fälle einen – dann hyperkorrekten – Schwund des *-d-* auf:

> CONSTRINGERE > *constrendre,* DEFENDERE > *defandre, def(f)andre,* *DIS-PENDERE > *despendre,* IN + DIRECTUS > *endroit, de-androit, (erendroit),* INTENDERE > *entendre,* PLANGERE > *plaindre, pleindre,* RENDERE > *rendre (-r-oit, -iens), randre* [aber: *randerois*], vendere > *vendre, vandre.*

Das gilt in gleicher Weise für eine Reihe von Eigennamen:

> *Alixandre(s), Andreu(s), Andriuns, Andryuns, Gendrenville, Gondricourt, Gundrecourt, Gondreville* und *Mandrez.*

Diese Stichprobe wurde vor der inzwischen abgeschlossenen, Dumitru Chihaï zu dankenden Detailkorrektur des Korpus durchgeführt; eine punktuelle Überprüfung der abweichenden Formen ergab denn auch in drei Fällen eine fehlerhafte Grundlage, einmal schon bei ARNOT *(Lanrecort* 128,29: ms. *Landrecort),* zweimal durch die Abschrift, die ganz intuitiv die Vorlage französisierte *(vendredi* 251,7: ARNOD *venredi*; *meprendre* 144,4: ARNOD *mepenre).* Allein dies zeigt die Gefahren, die in nicht ganz akribisch geprüften Korpora für die graphematische Auswertung schlummern.

Es blieben drei Urkunden mit unlothringischen Graphien (rechts die potentiellen Aussteller, also die im Text auftretenden Protagonisten):

charte 3	*semondront* 3, 28 vs. *tenr-* (bis)	comte de Bar vs. seigneur d'Apremont
charte 16	*prandre* 16, 63; 65 vs. *panre*, *tenr-*	prieuré de Flavigny vs. Saint-Vanne de Verdun; ducLorr, seigneur de Gondrecourt
charte 88	*prendre* 88,14 (bis) vs. *tenr-* (4x), *venr-*, *panre* (3x)	ducLorr vs. comte de Bar vs. comte du Luxembourg; Thibault, roi de Navarre

Bei den Urkunden 3 und 88 spricht die Verteilung der potentiellen Aussteller für die fortschrittliche Kanzlei der Grafen von Bar, bei der Urkunde 16 ist der Herzog von Lothringen ex negativo ausgeschlossen, da seine im Korpus gut vertretene Kanzlei ansonsten immer die lothringische Form wählt.

Die Gesamtverhältnisse entsprechen in etwa den von GOSSEN dargestellten Zahlen (GOSSEN 1967, 316: 97 % der lothringischen Urkunden *nr*, 3 % *ndr;* hier: durchweg *nr*, mit Ausnahme der Kanzlei der Grafen von Bar [ca. 5 % *ndr*] und der weiter nördlich liegenden Abtei Saint-Vanne de Verdun [von der das Priorat Flavigny abhing]). Aber die in diesem Beispiel zutage tretende überregionale Tendenz bei den Grafen von Bar bestätigt und verstärkt sich bei den anderen untersuchten graphematischen Eigenarten. Außerdem erbrachte bereits diese wenig ergiebige Untersuchung von *nr / ndr* Indizien für die Identifizierung dreier Aussteller; die Anwendung von zehn Kriterien konnte die Hälfte der etwa 160 nicht aufgund externer Elemente unmittelbar erkennbaren Aussteller faßbar machen.

6. Ausblick

Die hier zusammengestellten, stark kondensierten Bemerkungen beanspruchen keine Eigenwertigkeit und sollen zunächst einen Projektfortgang dokumentieren. Dennoch wird erkennbar, daß es mittels der originalen Dokumente ganz Nordfrankreichs möglich sein dürfte, erstmalig das Skriptoriennetz der Langue d'oïl zu eruieren. Damit entstünde ein wirklicher Anker, an dem alle makroskopischen und mikroskopischen Veränderungen in der Sprache zwischen 1200 und 1500 festgemacht werden könnten, auch für kopiale und literarische Texte.

Die Verknüpfung graphematischer und lexikometrischer Methoden wird weiterhin herausarbeiten können, wie groß die graphische oder lexikalische Varianz in bestimmten Skriptorien und in bestimmten Textsorten zu einem bestimmten Zeitpunkt war. So werden mikroskopische Veränderungen im Varietätenraum der Schrift nachweisbar, und die Entwicklung der französischen Schrift- und Standardsprache zwischen dem 13. und 15. Jahrhundert in Raum und Gesellschaftsstruktur kann nachgezeichnet werden. Insbesondere die Wege, auf denen sich Ausgleichs- und damit Standardisierungstendenzen durchgesetzt haben, sollten zutagetreten.

Es wird dann sogar möglich, Texte verschiedener Textsorten oder sogar verschiedener romanischer Sprachen auf einer abstrakten Ebene zu vergleichen, unter Beurteilung ihres Elaborationsgrads: Ist ein französischer Urkundentext der königlichen Kanzlei varianten- oder wortreicher als ein kastilischer oder toskanischer Urkundentext derselben Epoche oder als ein zeitgenössischer französischer Versroman?

Die großen Entwicklungen in der Sprachgeschichte der Romania sind uns bekannt. Was jetzt not tut, sind umfassende Detailuntersuchungen unter

Einsatz quantifizierender Methoden , die zwar ein hohes Zeitvolumen schon in der Vorbereitungsphase erfordern, dafür aber neue Wege erschließen können.

7. Zitierte Bibliographie

Arnod, Michel: Publication des plus anciennes chartes en langue vulgaire antérieures à 1265 conservées dans le département de Meurthe-et-Moselle. Thèse dactylographiée. Nancy 1974.

Bromm, Gudrun: Die Entwicklung der Großbuchstaben im Kontext hochmittelalterlicher Papsturkunden (elementa diplomatica 3). Marburg an der Lahn 1995.

Ernst, Gerhard / Wolf, Barbara: Textes français privés des XVII[e] et XVIII[e] siècles. CD-ROM-Ausgabe (Beihefte zur ZrP 310). CD 2. Tübingen 2002.

Gärtner, Kurt / Holtus, Günter / Rapp, Andrea / Völker, Harald (edd.): Skripta, Schreiblandschaften und Standardisierungstendenzen. Urkundensprachen im Grenzbereich von Germania und Romania im 13. und 14. Jahrhundert. Beiträge zum Kolloquium vom 16. bis 18. September 1998 in Trier (Trierer Historische Forschungen 47). Trier 2001.

Glessgen, Martin-Dietrich: Das altfranzösische Geschäftsschrifttum in Oberlothringen: Quellenlage und Deutungsansätze. In: Gärtner et al. 2001, 257–294.

Glessgen, Martin-Dietrich: L'élaboration philologique et l'étude lexicologique des 'Plus anciens documents linguistiques de la France' à l'aide de l'informatique. In: Frédéric Duval (ed.): Frédéric Godefroy. X[e] colloque international sur le moyen français (12–14 juin 2002, Metz). I. Dr. a

Glessgen, Martin-Dietrich: La lemmatisation: problèmes et méthodes. In: Pierre Kunstmann (ed.): Ancien et moyen français sur le Web: enjeux méthodologiques (4–5 octobre 2002, Ottawa). I. Dr. b

Gossen, Carl Th.: Französische Skriptastudien. Untersuchungen zu den nordfranzösischen Urkundensprachen des Mittelalters (Sitzungsberichte der Österr. Akad. der Wissenschaften. Phil.-hist. Kl., 253). Wien 1967.

Holtus, Günter / Völker, Harald: Editionskriterien in der Romanischen Philologie. In: Zeitschrift für romanische Philologie 115 (1999), 397–409.

Llamas Pombo, Elena: De Arte Punctandi. Antología de textos antiguos medievales y renacentistas. Salamanca 1999.

Möhren, Frankwalt: Édition et lexicographie. In: Martin-Dietrich Glessgen / Franz Lebsanft (edd.): Alte und neue Philologie (Beihefte zu editio 8). Tübingen 1997, 153–166.