

Abschlussbericht ADIUVARIS

Steffen Eger* Ineta Sejane†

19. April 2009

Inhaltsverzeichnis

1	Die Idee	2
2	Daten & der grundlegende Algorithmus	2
3	Aspekte und Probleme der Datenstrukturierung	5
4	Datenbank & Homepage	6
5	Ausblick	7
6	Tabellarische Auflistung der Tätigkeiten	7
7	Danksagungen	8
8	Anhang: Erläuterung linguistischer Begriffe	8

*eger.steffen@gmail.com

†sejane@ids-mannheim.de

1 Die Idee

Das IT-Projekt ADIUVARIS hat sich die Realisierung einer Internetplattform für das Lateinische zum Ziel gesetzt. Allgemein gesprochen behandelt ADIUVARIS dabei Problemstellungen auf drei Ebenen: auf der lexikalischen, der grammatischen und der historisch-kulturellen Ebene. Auf der **lexikalischen Ebene** soll es vor allem darum gehen, lexikalische lateinische Einheiten mit lexikalischen Einheiten der deutschen Sprache in Beziehung zu setzen, mit anderen Worten: ADIUVARIS bietet Übersetzungen aus dem Lateinischen und ins Lateinische an. Auf der **grammatikalischen Ebene** soll es darum gehen, eine vollständige grammatische Analyse lateinischer Vollformen anzubieten, d.h. neben einer ‘groben’ Kategorisierung lateinischer Vollformen in Wortarten (Verb, Substantiv, Adjektiv, Adverb, etc.) soll auch eine ‘feinere’ Unterkategorisierung (1. Person Präsens, etc.) vorgenommen werden. Dieses Problem hat zwei unterschiedliche Aspekte, die sich in ihrem Schwierigkeitsgrad — aus Sicht der automatisierten Verarbeitung — stark unterscheiden. Einerseits können Einzelwörter ohne Kontext betrachtet werden, in welchem Fall in der Regel Ambiguität, d.h. Mehrdeutigkeit, auftritt, da oft ein und dieselbe Wortform mehrere Analysen zulässt (bspw. lässt die Form *insula* (Insel) in der Analyse sowohl den Kasus ‘Nominativ’ als auch ‘Ablativ’ zu). Diese Ambiguität lässt sich ohne weiteren Kontext nicht auflösen und wird dem Benutzer in dieser Form zur Verfügung gestellt. Andererseits ist die Zuordnung von grammatischer Analyse zur Wortform im Satzkontext (in den allermeisten Fällen) eindeutig, da der Satz unter den verschiedenen Lesarten disambiguiert (vgl. *Marcus in insula vivebat* (Ablativ) vs. *Italia non insula est* (Nominativ)). Diese Zuordnung stellt für die automatisierte Sprachverarbeitung ein nicht-triviales Problem dar, insbesondere, da Latein als Paradebeispiel einer synthetischen Sprache viele Eigenschaften aufweist (insbesondere freie Wortstellung, etc.), die die Lösung des Problems erschweren.

Schließlich soll es auf einer **historisch-kulturellen Ebene** darum gehen, Informationen zur lateinischen Sprache und Kultur bereitzustellen, möglicherweise in Form eines Forums, das Gedankenaustausch anregt, und in der Form von Zitatensammlungen und Sprichwortsammlungen.

ADIUVARIS richtet sich an Schüler und Studenten des Lateinischen, hat also eine stark praxis-orientierte Zielsetzung, was sich u.a. in der Festlegung dessen, was das ‘Lateinische’ umfasst (klassisches Latein mit besonderem Gewicht auf Cicero, Caesar, etc.), niederschlägt. Kostenfrei soll den Schülern ein ständig über das Internet verfügbares Instrumentarium zur Verfügung gestellt werden, mit dessen Hilfe der Zugang zu lateinischen Texten wesentlich erleichtert wird.

2 Daten & der grundlegende Algorithmus

Ausgangspunkt eines ersten Teils des Projekts, der sowohl die oben beschriebene lexikalische wie Teile der grammatischen Ebene umfasst, ist ein in XML verfasstes Wörter-

buch, dessen Einträge neben einer lateinischen Grundform¹ auch grammatische Angaben (z.B. Geschlecht, Flexionsklasse, etc.) sowie eine deutsche Übersetzung beinhalten. Exemplarisch sei hier der Wörterbucheintrag der Grundform “*amo, amavi, amatum, amare*” dargestellt.

```
<entry POS="verb" STATUS="regular" INFLECTION_CLASS="A-Conjugation">
  <translation LANGUAGE="german">lieben , verliebt sein</translation>
  <form PERSON="first" NUMBER="sg" MOOD="indicative"
        TENSE="present" VOICE="active">amo</form>
  <form PERSON="first" NUMBER="sg" MOOD="indicative"
        TENSE="perfect" VOICE="active">amavi</form>
  <form SPECIAL="PPP">amatum</form>
  <form MOOD="infinitive" TENSE="present" VOICE="active">amare</form>
</entry>
```

Während *irreguläre* Verben, Substantive und Adjektive, d.h. solche, deren Flexion einem nicht-regulären Muster folgt, direkt aus dem XML-Wörterbuch in eine Datenbank übertragen werden, ist die Vorgehensweise eines von uns entworfenen Algorithmus bei *regulären* Vokabeln hingegen so, dass aus einer gegebenen Grundform-Angabe sowie der grammatischen Information sowohl ein Stamm als auch seine Endungen bestimmt werden. Beispielsweise wird aus der substantivischen Grundform “*vinum, vini*” (Wein) der Stamm *vin-* extrahiert, an den dann die Endungen der O-Deklination, *-um, -i, -o, -um, -o, -um, -a, -orum, -is, -a, -is, -a*, angehängt werden. Aus Stamm und Endungen werden Vollformen erzeugt, die anschließend in der Datenbank abgespeichert werden. Erwähnenswert ist dabei, dass einem durchschnittlichen regulären Substantiv 12 Vollformen, einem Adjektiv bis zu 100, einem Verb gar bis zu 240 Vollformen zugeordnet sind (vgl. Tabelle 1). Stellvertretend zeigen wir das Input/Output-Verhalten des Algorithmus bei Eingabe der Grundform “*amo, amavi, amatum, amare*”:

```
Dictionary dict = new Dictionary ();
Inflector inflector = new Inflector (dict );
inflector.setForms ("amo" , "amavi" , "amatum" , "amare" );
inflector.setTranslation ("lieben , verliebt sein" );
inflector.setInflectionClass ("a-conjugation" );
inflector.inflect ();
```

Hierfür liefert der Algorithmus auszugsweise den folgenden Output:

```
amo,v1spia---,1
amas,v2spia---,1
amat,v3spia---,1
amamus,v1ppia---,1
amatis,v2ppia---,1
```

¹Bspw. ist das bei lateinischen Substantiven normalerweise der Nominativ Singular und der Genetiv Singular, bei Verben ist es in der Regel die 1. Person Präsens Singular, der Perfekt, das Partizip Perfekt und der Infinitiv.

amant,v3ppia---,1

wobei die erste Spalte die jeweilige Form angibt, die zweite ihre grammatische Spezifikation (“verb,1.person,singular,present,indicative,active”, etc.) und die dritte auf die Grundform “*amo, amavi, amatum, amare*” verweist. Diese Ausgabe wird dann in die Datenbank eingefügt.

Wortklasse	#Grundformen (regulär)	#Grundformen (irregulär)	#Vollformen
Verb	593	53	141917
Substantiv	681	11	8261
Adjektiv	318	1	24530
Adverb	155		166
Konjunktion	51		51
Präposition	29		29
Partikel	8		8
Numeral	151		2358
Pronomen	118		1459
Phrase	16		16
Rest	1		1
	2121	65	178796

Tabelle 1: Verteilung der Grundformen und Vollformen in unserem Wörterbuch.

In einem zweiten Teil des Projekts geht es um die in Kapitel 1 beschriebene Analyse lateinischer Vollformen im Satzkontext, d.h. um die jeweilige grammatische Beschreibung aller lateinischen Worteinheiten in einem gegebenen Satz. Wie bereits angesprochen, ist dies ein nicht-triviales Problem, für das in der Literatur Lösungsansätze vornehmlich für analytische Sprachen wie das Englische konzipiert worden sind. Unser Ziel war es, neue Lösungsansätze für synthetische Sprachen aufzuspüren, da das Standard-Lösungs-Modell, das Hidden-Markov-Modell, nur sehr lokale Abhängigkeiten beschreiben kann. Aufgrund der relativ freien Wortstellung müssen im Lateinischen aber globalere Abhängigkeitsstrukturen betrachtet werden. Zwar können wir in diesem Bereich noch keinen Durchbruch vermelden, da verschiedene Teile des Projekts einen größeren Zeitaufwand gefordert haben als ursprünglich gehofft, doch haben wir bereits eine klare Vorstellung über das weitere Vorgehen: Ordnet man nämlich jedem einzelnen lateinischen Wort in einem Satz sämtliche (kontext-unabhängigen) möglichen Analysen zu — und aufgrund des oben beschriebenen Algorithmus² können wir das — so ergibt sich für den gesamten Satz eine Menge von Lösungskandidaten², unter denen wir den ‘wahrscheinlichsten’ Anwärter aussuchen wollen. Es gibt nun im Perseus-Projekt

²Wobei ein ‘Lösungskandidat’ eine Folge von grammatischen Labels ist, die den Wörtern des Satzes zugeordnet wird, z.B. wäre (Substantiv Nominativ, Präposition, Substantiv Ablativ, Verb Imperfekt) ein möglicher Kandidat für den Satz *Marcus in insula vivebat*.

(<http://nlp.perseus.tufts.edu/syntax/treebank/>) Daten, die es uns ermöglichen, statistische Berechnungen über die Verteilung lateinischer grammatischer Labels (*tags*) zu machen. Weiterhin erlaubt der aus der künstlichen Intelligenz bekannte *decision list*-Algorithmus die Inbetrachtung eines beliebig langen Kontexts, umgeht also die Schwäche der Lokalität des Hidden-Markov-Modells.

Wir sind zuversichtlich, durch die Verbindung beider Komponenten — der Daten aus dem Perseus-Projekt und dem genannten *decision list*-Algorithmus — innerhalb der nächsten beiden Monate sehr brauchbare Resultate zu präsentieren, die wir dann publizieren können, da es sich hierbei, wie bereits erwähnt, um Neuland in der Forschung handelt.

3 Aspekte und Probleme der Datenstrukturierung

Die Akquirierung und Strukturierung der Ausgangsdaten für ADIUVARIS erwies sich als schwieriger und zeitaufwändiger als zunächst vermutet. So stand uns zwar ein rudimentäres **Wörterbuch** zur Verfügung, das von der Altphilologie der Universität Heidelberg bereitgestellt wurde. Dieses war einerseits aber in seinem Umfang und seiner Abdeckung sehr beschränkt, sodass manuelle Ergänzungen unumgänglich waren. Andererseits war dieses Ausgangswörterbuch in einer Form verfasst, die für die automatisierte Verarbeitung ungeeignet war. So wurden beispielsweise viele Abkürzungen (*vinum, -i, n*) verwendet, ebenso war der Gebrauch der Bezeichnungen der Flexionsklassen uneinheitlich, und die Übersetzungen enthielten vielfältige Kommentare, die sich auf jeden Aspekt eines Wörterbuch-Eintrags (Grammatik, lateinische Form, deutsche Form) beziehen konnten. Hier war also händische Nachbearbeitung und Korrekturlesen in nicht unerheblichem Maße vonnöten. Schließlich entschlossen wir uns, das ursprüngliche CSV-Format aufzugeben, da sich unregelmäßige lateinische Wortformen (*esse, ferre, etc.*) damit nur schwer darstellen ließen, weshalb wir uns für das flexiblere XML entschieden. Darüber hinaus bereiteten uns die lateinischen **Adjektive** größere Probleme, da bei ihnen nicht nur alle drei Genera (männlich, weiblich, sächlich) berücksichtigt werden müssen, sondern sich die Realisierung dieser Genera auch von Adjektiv zu Adjektiv unterscheidet (bei ein-endigen Adjektiven haben alle Genera dieselbe Realisierung, bei zwei-endigen haben zwei der drei Genera dieselbe Realisierung und bei drei-endigen hat jedes Geschlecht seine eigene Formstruktur³). Weiterhin bilden manche Adjektive einen Komparativ und Superlativ, andere hingegen nicht, ebenso kann man nur manchen Adjektiven ein Adverb zuordnen. Diese Informationen waren nicht in unserem Ausgangswörterbuch enthalten und ließen sich ohne sehr genaue Kenntnis der lateinischen Sprache nur mit erheblichem Aufwand produzieren, sodass wir die Hilfe eines Experten zuzogen. Darüber hinaus war auch nicht immer klar, welche und wieviele **Wortklassen** verwendet werden sollen, da in vielen Fällen eine eindeutige Zuordnung entweder schwer fiel oder gar nicht möglich war (Welcher Wortklasse gehört *semi-* (halb) an? Können Konjunktionen immer eindeutig von Partikeln und

³Bspw. lauten die Nominative für die Adjektive *scharf, gleich* und *abwesend*: *acer (m), acris (f), acre (n)* (drei-endiges Adjektiv), *aequalis (m/f), aequale (n)* (zwei-endig), und *absens (m/f/n)* (ein-endig).

Präpositionen unterschieden werden? Was ist der Status von Zahlwörtern und Pronomina?) Es muss noch erwähnt werden, dass die wichtigsten Wortklassen Subklassifizierungen zulassen (bei Substantiven: plurale tantum, singulare tantum, bei Verben: Deponentien, Semideponentien, ohne Perfektstamm), die bei der Repräsentation der Daten und beim Verhalten der Algorithmen berücksichtigt werden müssen. Schließlich musste die einer lateinischen Grundform zugeordnete Menge von deutschen **Übersetzungen** in einzelne Einheiten aufgetrennt werden, um effiziente Abfragen Deutsch-Latein zu gewährleisten. Auch dieses Unterfangen ließ sich aufgrund der oben genannten Probleme nur teilweise automatisiert durchführen.

4 Datenbank & Homepage

ADIUVARIS arbeitet mit einer MySQL-Datenbank. In diese werden alle von den Algorithmen erzeugten Daten eingetragen. Wir zeigen eine verkürzte Hierarchie der Tabellen (siehe Abbildung 1).

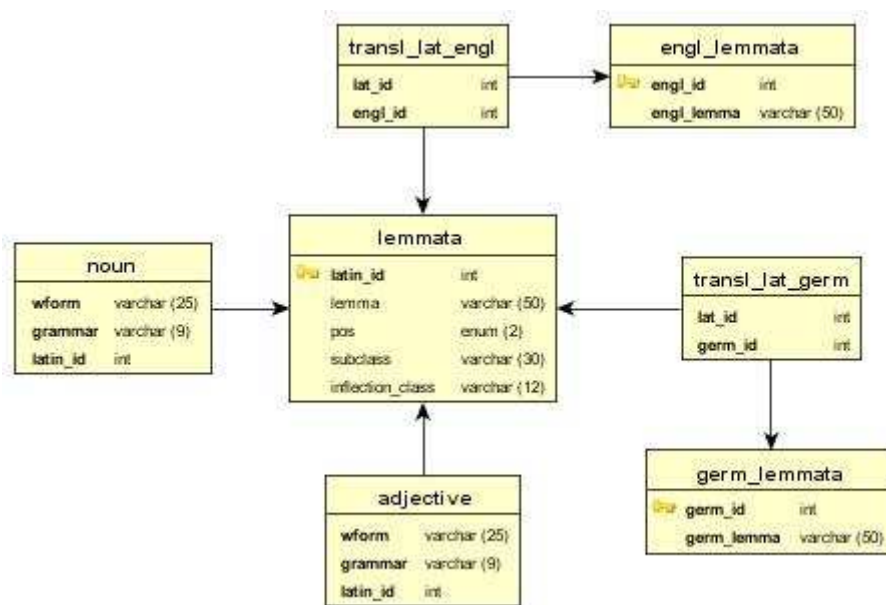


Abbildung 1: Tabellenstruktur unserer Datenbank, ausgewählte Tabellen. Die Tabelle ‘Lemmata’ fasst die Grundformen. Die Tabellenstruktur sieht bereits die Möglichkeit vor, weitere Sprachen (hier: Englisch) in das System einzubeziehen.

ADIUVARIS hat ebenfalls eine Homepage, die auf die Datenbank mittels PHP zugreift. Die Homepage ist unter <http://www.adiuvaris.de> abrufbar, reserviert ist ebenfalls die Domäne <http://www.adiuvaris.com>. Wir hoffen, die Homepage später noch durch Werbung finanziell nutzbar machen zu können. Zum gegenwärtigen Zeitpunkt erlaubt die Homepage Abfragen in den Richtungen Deutsch-Latein und Latein-Deutsch, sie stellt

zudem grammatische Analysen für aktuell ca. 178.000 Vollformen (siehe oben) bereit. Komplexe grammatische Suchanfragen sind möglich.

5 Ausblick

Es gibt eine Menge von Problemen, die von ADIUVARIS in der nächsten Zeit noch angegangen werden müssen, um das Projekt zu einem erfolgreichen Abschluss zu führen. Die folgende Aufzählung benennt diese kurz.

- Grammatische Analyse im Satz. Wie oben beschrieben, glauben wir, dass wir dieses Problem bald ‘gelöst’ haben werden. Falls das passiert, haben wir nicht nur zu aktueller Forschung beigetragen, sondern bieten dann auch einen Service an, den es für das Lateinische so bisher nirgends gibt.
- Einbindung weiterer Sprachen. Als erstes wollen wir vor allem das Englische einbinden, um dann die Sprachpaare Latein-Deutsch und Latein-Englisch anbieten zu können. Damit hätte unser Projekt auch internationale Präsenz.
- Die historisch-kulturelle Ebene konnte von ADIUVARIS bisher nicht abgedeckt werden. In einem nächsten Schritt wollen wir unseren Usern ein Forum anbieten, in dem sie sich austauschen können. Ebenso soll eine lateinische Sprichwortsammlung online gestellt werden. Weiterhin sollen die User aktiv an ADIUVARIS teilnehmen können, d.h. eigene Wörterbucheinträge liefern, etc. Davon versprechen wir uns eine Explosion unserer anzubietenden Datengrundlage (Stichwort Web 2.0 bzw. Mitmach-Internet). Außerdem könnte das Design der Webseite attraktiver und benutzergruppen-gerechter gestaltet werden.
- Bisher arbeiten wir noch mit einer alten Datenbank-Hierarchisierung, bei der die Zugriffe besonders im Bereich Deutsch-Latein zu langsam bearbeitet werden. Wir wollen bald endgültig auf unsere neue, oben geschilderte Strukturierung umsteigen.
- Die deutschen Übersetzungen müssen qualitativ und quantitativ verbessert werden. Auch muss eine verbesserte Lösung für die Repräsentation grammatischer Zusatzinformationen, die z.B. das Ko-okkurrenzverhalten einzelner Worteinheiten beschreiben (z.B. *abstinere* (+Abl.), AcI, NcI), gefunden werden.

6 Tabellarische Auflistung der Tätigkeiten

Wir geben eine kurze Auflistung unserer Tätigkeiten. Die Ermittlung dieser Auflistung ist vor allem deshalb schwierig, weil viele Tätigkeiten unterbrochen wurden und erst nach einiger Zeit fortgesetzt werden konnten bzw. weil neue Konzeptionalisierungen Überarbeitungen erforderlich machten. Einige Aufgaben erforderten eine andauernde, wenn auch nur eingeschränkte Aktivität.

Tätigkeit	Dauer (in Monaten)	Zeitpunkte
Algorithmus (Vollformen)	3	April, Mai, August
Wörterbuch-Überarbeitung	6	Mai, September, Oktober, November, Dezember, März
Homepage	3	Juni, Juli, August
Datenbank (Erstellung, Strukturierung)	6	Mai, Juni, Dezember Januar, Februar, März
Algorithmus (Satzebene)	3	Dezember, Januar, Februar

7 Danksagungen

Wir danken denen, die uns geholfen haben: Christian Jäkel, Kasia Krysinska, PD Dr. Wolfgang Merkle, Hendrik Niederlich. Wir sind auch dem Perseus-Projekt für die Bereitstellung lateinischer Texte und deren Annotation zu großem Dank verpflichtet.

Besonderen Dank an die MFG-Stiftung für die finanzielle Unterstützung und fördernde Betreuung, die so vieles erst möglich gemacht haben.

8 Anhang: Erläuterung linguistischer Begriffe

Flexion bezeichnet in der Grammatik die Änderung der Gestalt eines Wortes zum Ausdruck seiner grammatischen Funktion im Satz (Bsp. *das Haus, des Hauses, der Häuser*, etc.), es wird abhängig von der Wortart unterschieden nach der Flexion bei Substantiven und Adjektiven (**Deklination**) und der Flexion bei Verben (**Konjugation**). Adjektive und Adverbien bilden außerdem einen **Komparativ** (*schneller*) und **Superlativ** (*am schnellsten*).

Genus (Pl. Genera) bezeichnet das (**grammatische**) **Geschlecht** (männlich, weiblich, sächlich) eines Substantivs.