Hausarbeit

zur Erlangung des Magistergrades
an der Ludwig-Maximilians-Universität München

Erstellen einer Lateingrammatik im Grammatical Framework

vorgelegt von Herbert Lange

Fach: Computerlinguistik

Referent: Prof. Dr. Klaus U. Schulz

München, den 30.9.2013

Inhaltsverzeichnis

1 Einleitung

1.1 Motivation

So mancher, der den Titel dieser Arbeit liest, wird sich wundern, warum man in der heutigen Zeit eine computergestützte Grammatik gerade für eine tote Sprache wie Latein entwickeln will. Doch die konkrete Sprache, die umgesetzt werden sollte, war bei der Wahl des Themas zunächst zweitrangig. Die Intention hinter dieser Arbeit war es eher, einmal in einem konkreten Falle die im Studium behandelten Theorien der Morphologie und der Syntax, aber auch die Prinzipien der Lexikonerstellung, in einem einheitlichen Projekt zusammenzuführen.
Das fuer dieses Unterfangen am ehesten geeignete Softwaresystem schien schon sehr bald das Grammatical Framework[1] zu sein. Es stellt alle benötigten Hilfsmittel zur Verfügung, die jeweils für die einzelnen Komponenten benötigt werden, sorgt aber auch durch einen einheitlichen Beschreibungsformalismus für die nötige Konsistenz zwischen allen Bestandteilen. Weitere Vorteile des Grammatical Frameworks sind der mächtige Beschreibungsformalismus für Grammatiken, Unterstützung für Multilingualität und aktive Entwicklung als Open Source-Software.
Nachdem sich das Grammatical Framework als geeignet heraus gestellt hatte, fiel die Wahl der zu bearbeitenden Sprache auf Latein, da diese Sprache, die trotz ihres Alters in der Linguistik weiterhin nicht unbedeutend ist, in der Ressource Grammar Library[2] bisher nur sehr rudimentär umgesetzt war.

1.2 Inhalt

Im Folgenden sollen zunächst die Grundlagen der Arbeit genauer geschildert werden, es folgt also eine genauere Betrachtung des Grammatical Framework so wie der lateinischen Sprache. Anschließend wird das Vorgehen bei der Implementierung der Grammatik als zukünftiger Bestandteil der Ressource Grammar Library geschildert werden. Und zum Schluss soll noch eine Betrachtung der Erweiterungs- und Anwendungsmöglichkeiten folgen.

1.3 Das Grammatical Framework

Das Grammatical Framework ist ein Softwaresystem mit einer spezialisierten Programmiersprache um Grammatiken zu entwickeln. Es bietet die nötigen Möglichkeiten um natürliche Sprachen zu verarbeiten. Dabei benutzt es Formalismen, wie sie auch in modernen funktionalen Programmiersprachen wie Haskell zu finden sind.[3]

[1] http://www.grammaticalframework.org/
[2] http://www.grammaticalframework.org/lib/doc/synopsis.html
[3] RANTA S. vii

S
⇒ NP VP
⇒ Det N VP
⇒ “der” N VP
⇒ “der Mann” VP
⇒ “der Mann” V
⇒ “der Mann schläft”

Abbildung 1: Ableitung eines Satzes

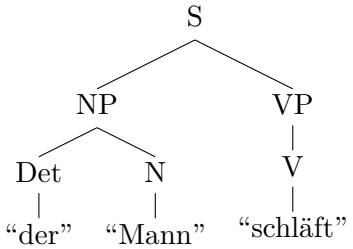

Abbildung 2: Entsprechender Syntaxbaum

Die große Stärke dabei ist die Multilingualität. Grundkonzept dabei ist die Trennung in eine konkrete und eine abstrakte Repräsentation der Grammatik. Dabei ist die konkrete repräsentation jeder Sprache eigen während die abstrakte Repräsentation von mehreren Sprachen geteilt werden kann. Über diesen Schritt der abstrakten Repräsentation kann man eine Übersetzung zwischen allen Sprachen umsetzen, die diese abstrakte Syntax teilen.[4] Im folgenden soll darauf noch genauer eingegangen werden.

1.3.1 Der Grammatikformalismus

Meist werden im Bereich der Computerlinguistik und Informatik kontextfreie Grammatiken, also Grammatiken von Typ 2 der Chomsky-Hierarchie verwendet.[5] Dies hat meist den Grund, dass die Ausdrucksmächtigkeit größtenteils ausreicht, jedoch der Verarbeitungsaufwand vergleichsweise gering ist.[6] Die in Beispiel 1 gegebene Grammatik ist

$S \longrightarrow NP\ VP$
$NP \longrightarrow Det\ N$
$N \longrightarrow$ Mann
$Det \longrightarrow$ der
$VP \longrightarrow V$
$V \longrightarrow$ schläft

Beispiel 1: Kontextfreie Grammatikfragment

ein sehr minimalistisches Beispiel für eine kontextfreie Grammatik. Mit ihrer Hilfe kann nur der eine deutsche Satz *Der Mann schläft* hergeleitet werden. Dabei hat die Ableitung die in Beispiel 1 gezeigte Form. Im Formalismus des Grammatical Framework wird die oben gegebene Grammatik in die abstrakte und die konkrete Syntax zerlegt. Dabei entspricht die abstrakte Syntax dem Syntaxbaum ohne die terminalen Blätter.

[4] RANTA S. 10ff.
[5] quelle
[6] quelle

```
abstract Satz = {
  flags startcat = S ;
  cat S ; NP ; VP ; Det ; N ; V ;
  lin
    mkNP : Det -> N -> NP ;
    mkVP : V -> VP ;
    mkS : NP -> VP -> S ;
    der_Det : Det ;
    Mann_N : N ;
    schlafen_V : V ;
}
```

Beispiel 2: Abstrakte Syntax

```
concrete Satz of SatzAbs = {
  lincat S,NP,VP,Det,N,V = Str;
  lin
    mkNP det n = det ++ n ;
    mkVP v = v ;
    mkS np vp = np ++ vp ;
    der_Det = "der" ;
    Mann_N = "Mann" ;
    schlafen_V = "schlaeft" ;
}
```

Beispiel 3: Konkrete Syntax

1.3.2 Die Ressource Grammar Library

Was für allgemeine Programmiersprachen eine Standardbibliothek ist, ist im Grammatical Framework für die Multilingualität die Ressource Grammar Library. Sie ist definiert als gemeinsame abstrakte Syntax, die für verschiedenen Sprachen implementiert ist. Auf diese Möglichkeit ist eine grundlegende Übersetzung zwischen den unterstützten Sprachen direkt nach der Installation möglich. Meist muss jedoch mindestens das nötige Vokabular angegeben werden, da das Lexikon auf eine kleine Anzahl von Wörtern beschränkt ist, die benötigt wird um die grammatischen Konstrukte zu testen.

1.4 Die Lateinische Sprache

1.4.1 Sprachwissenschaftliche Einordnung

Die lateinische Sprache gehört zur indogermanische Sprachfamilie und dort zur Unterfamilie der italischen Sprachen. Entstanden ist es als ein in der Stadt Rom üblicher Dialekt parallel zu weiteren ländlichen Dialekten im Latium, im laufe der Zeit verdrängte es jedoch die weiteren italischen Sprachen im Zuge der Ausdehnung des römischen Reichs.[7] Die Sprachgeschichte kann in mehrere Epochen unterteilt werden, nämlich das Altlatein, das klassische Latei, das Mittellatein (ca. 650 n. Chr. bis ca. 1400 n. Chr.) und das Neulatein (ca. 1400 n. Chr. bis heute).[8] Auch heute noch am bedeutendsten ist wohl das klassische Latein, das weiterhin in Schulen unterrichtet wird und vor allem mit seinem großen überlieferten Textkorpus hervorsticht.
Latein gehört zu den stark flektierenden Sprachen. Es gibt fünf zum teil genusbasierte Flektionsklassen für Nomen, sechs Verschiedene Kasus (Nominativ, Genitiv, Dativ, Akkusativ,

[7] METZLER2004 S. 5359
[8] MÜLLER-LANCE2006 S. 27ff.

Ablativ und Vokativ), drei Genera (Maskulin, Feminin, Neutrum), ein voll flektierendes Pronomensystem und vier relativ stark synthetische Deklinationsklassen für Verben.[9] Zu den Kasus sei anzumerken, dass der Ablativ ein eigenständiger Kasus ist, jedoch der Vokativ oft mit dem Nominativ zusammenfällt.[10]
Die Wortstellung des Lateinischen wird oft als sehr frei beschrieben, allerdings gibt es eine klare Präferenz der SOV-Wortstellung im Satz, also dass das Objekt des Satzes direkt auf das Subjekt folgt, und das Verb den Satz abschließt. Die position des Adjektivs im Bezug auf das Nomen ist allerdings wirklich recht frei.[11]

1.4.2 Bedeutung in der heutigen Zeit

Man kann sich natürlich über die Notwendigkeit streiten, sich in der heutigen Zeit noch mit der lateinischen Sprache zu beschäftigen. Es gibt aber auch ziemlich gute Gründe dafür Latein nicht einfach nur als tote Sprache abzustempeln und nicht weiter zu betrachten.
So gibt es verschiedenste Personengruppen, für die Lateinkenntnis von Vorteil ist.

2 Grammatikerstellung

2.1 Lexikon

Den Beginn dieser Grammatikimplementierung bildete die Erstellung des minimal nötigen Lexikons. Durch die abstrakte Syntax der RGL[12] eine Liste von ca. 400 englischen Bezeichnern für Worte vorgegeben, die in jeder Sprache umgesetzt werden sollten.
Um für das vorgegebene Vokabular die passenden lateinischen Entsprechungen zu finden, wurde verschiedene Vorgehensweisen angewandt.
Für die meisten englischen Begriffe war es zunächst problemlos möglich, deutsche Entsprechungen zu finden. Bei problematischeren Begriffen wurde ein verbreitetes Onlinewörterbuch[13] zu Rate gezogen. Somit war es für fast alle vorgegebenen Begriffe möglich, eine adequate deutsche Übersetzung zu finden. Die einzige Art von Wörtern, die weiterhin zu Problemen führten, waren Wörter mit ambiger Bedeutung, wie das häufig gezeigte Wort *bank*, das in vielen Sprachen mehrer verschiedene Bedeutungen haben kann, z.B. im Deutschen als Sitzgelegenheit und als Geldinstitut oder im Englischen ebenfalls als Geldinstitut oder als Flussufer.[14] Für diesen und ähnliche Begriffe wurde willkürlich eine plausible Bedeutung gewählt, da keine Hinweise zur gewünschten Bedeutung in der Grammar Library gefunden werden konnte. Die Entscheidung eine einzige Bedeutung zu wählen, und nicht verschiedene Bedeutungen als Varianten des Wortes zu implementieren, wurde getroffen

[9] METZLER2004 S. 5359
[10] ???
[11] METZLER2004 s. 5359
[12] vgl. lib/src/abstract/Lexicon.gf
[13] http://dict.leo.org
[14]

um die Anzahl der möglichen Übersetzungen möglichst gering zu halten.
Nachdem für alle Bezeichner im abstrakten Lexikon eine zwischenzeitliche deutsche Entsprechung, nach dem obigen Schema, gefunden wurde, wurde versucht, diese deutschen Begriffe in die lateinische Sprache zu übersetzen. Dies geschah zum größten Teil mit Hilfe des deutsch-lateinischen Teils des Standardwörterbuchs[15], soweit ein entsprechender Eintrag im diesem Wörterbuch zu finden war. Zusätzlich zu den recht kurzen Einträgen in diesem Teil des Wörterbuch, wurden auch alle weiteren vefügbaren Informationen zu den gefundenen lateinischen Begriffen berücksichtigt.
Bei vielen, meist moderneren Begriffen, konnte nicht immer ein entsprechender Wörterbucheintrag gefunden werden. Wenn auch in anderen verfügbaren Wörterbüchern[16] kein Eintrag zu finden war, gab es noch die Möglichkeit, auf Internetquellen zurückzugreifen, die meist auf dem Prinzip der freiwilligen Kollaboration basieren. Eine der interessantes Quelle für moderne Begriffe aus dem Breich der Substantive ist wohl die lateinische Wikipedia[17]. Obwohl Latein als tote Sprache gilt, existieren dort über 90000 lateinische Artikel[18]. Natürlich muss man immer bedenken, dass es keine Garantie für die Qualität von kollaborativen Onlinequellen gibt.

[15] Langenscheidt
[16] PONS
[17] http://la.wikipedia.org/wiki/Pagina_prima
[18] http://la.wikipedia.org/wiki/Specialis:Census; Stand: 30.7.2013