Ludwig-Maximilians-Universität München
Fakultät für Sprach- und Literaturwissenschaften
Centrum für Informations- und Sprachverarbeitung
Wintersemester 2013/2014

Vorlesung: *Symbolische Programmiersprache*

Guidelines für das automatische und manuelle POS-Tagging der armenischen Sprache

Dozentin: *Desislava Zhekova*

vorgelegt von:

Tetsuhiro Tamada
Matrikelnummer: 10987470
E-Mail: tetsuhirotamada@gmail.com
Tsovinar Sekhleyan
Matrikelnummer: 10592671
E-Mail: tsovsekh@gmail.com

Inhaltsverzeichnis

1. Einleitung

Im Bereich der maschinellen Textverarbeitung wurden viele Vorschritte in verschiedenen Sprachen gemacht. Eine von diesen Bereichen ist die Wortartenerkennung, sogenannte Part-of-speech Tagging, die zu verschiedenen linguistischen Zwecken der maschinellen Textverarbeitung verwendet werden kann. Die vorliegende Arbeit stellt hauptsächlich die Anleitung der manuellen und der automatischen Wortartenerkennung der armenischen Sprache dar, in der bis jetzt noch keine Forschungen gemacht wurden.

Zunächst wird in unserer Arbeit für die armenische Sprache erstelltes POS-Tagset präsentiert, das alle Wortarten mit einigen flektierten Wortformen, wie Numerus, Tempus, Modus umfasst. Anhand eines Textes (in dieser Arbeit „Corpus“ genannt) von Wikipedia wird jedem Wort sein entsprechender Tag („Tagged_Corpus“) zugewiesen.

-Beispiel-

Ժամանակակից/JJ Հայաստանի/NP Հանրապետությունը/NP…(NN = Nomen, JJ = Adjektiv, NP = Eigenname) – *Die jetzige Armenische Republik …*

Einige Namen der Tags wurden von dem Penn Treebank Projekt genommen und einige von den ganzen englischen Namen der Wortarten verkürzt.

In zweiten und dritten Kapiteln wird beschrieben, was bei der automatischen Segmentierung und Wortartenerkennung der armenischen Texte zu beachten ist. Für diesen Zweck wurden zwei Programme „Tokenizer“ und „Tagger“ entwickelt und bei „Corpus“ geprüft und verbessert.

Das vierte Kapitel enthält die ganze POS-Tagset Liste der entsprechenden Wortarten mit ihren Wortformen. Im fünften Kapitel werden die einzelnen Tags genauer mit Beispielen beschrieben. Abschließend werden einige Beispiele der Disambiguierung der Wörter gebracht.

Die kompletten Programme “Tokenizer“ und „Tagger“ mit der Textdatei „Corpus“, „Tagged_Corpus“ sind unter URL: https://github.com/ttamada/ArmenianPOSTagging einzusehen.

2. Tokenisierung der armenischen Sprache

Die Tokenisierung der armenischen Sprache erfolgt wie im Deutschen. Der Satz wird an Punktuationszeichen und Leerzeichen getrennt. Es sind einige Ausnahmen bei der Tokenisierung zur Beachtung, die in einem weiteren Teil dieses Kapitels beschrieben werden.

-Beispiel-

Աննան, Յուլիանը և Կարինան դասընկերներ են:

Anna, Julian und Karina Schulfreunde sind.

Tokenisiert: (Աննան) (,) (Յուլիանը) (և) (Կարինան) (դասընկերներ) (են) (:)

2.1. Das Verfahren der Tokenisierung

Trennung an Punktuationszeichen und Leerzeichen- Die armenische Sprache hat sowohl eigene als auch international anerkannte Punktuationszeichen. Die armenischen Punktuationszeichen sind: Satzendung (:), dem Komma gleichgetellt (.), armenisches Anführungszeichen («»), Betonungszeichen (՛),dem Komma gleichgetellt (՝), Ausrufezeichen (~), Fragezeichen(՞), Zeichen für den unvollständigen Satz (…), Klammern ([], ()). Andere Satzzeichen sind z.B. %$& usw.

-Beispiel-

Text: Այն փոքրիկ տղան՝ իմ եղբայրը… - *Der kleine Junge, mein Bruder…*

Tokenisiert: (Այն) (փոքրիկ) (տղան) (՝) (իմ) (եղբայրը) (…)

Zwei Zeichen davon, das Fragezeichen und Betonungszeichen können entweder auf dem letzten Vokal oder am Ende des Wortes vorkommen.

Beispiel:

- Text: Քանիսի՞ն կսկվի դասը: - *Um wie viel Uhr fängt der Unterricht an?*

  Tokenisiert: (Քանիսին) (՞) (կսկվի) (դասը) (:)

- Text: Խոսի՛ր ինձ հետ: - *Rede mit mir!*

  Tokenisiert: (Խոսիր) (՛) (ինձ) (հետ) (:)

Zur Ausnahme der Tokenisierung gehören einige Abkürzungen, deren Punkte bei der Tokenisierung nicht getrennt werden sollen.

-Beispiel-

Text: Մ.թ. 301 թ.-ին…- *im Jahre 301 n. Chr….*

Լուիզան ասաց.-Գնանք տուն:-*Luiza sagte: „Lass uns nach Hause gehen. “*

Tokenisiert: (Մ.թ.) (301) (թ.-ին) (…)

(Լուիզան) (ասաց) (.) (-) (Գնանք տուն) (:)

Zur Ausnahme der Tokenisierung gehören auch die Mehrwortlexeme, die für weitere Bearbeitung dieser Wörter vom Tokenizer als Ganzes angesehen werden sollen.

Beispiel:

- Text: **Ինչ-որ բան** խանգարում էր իրեն: - ***Etwas*** *hat ihn gestört.*

  Falsch tokenisiert: (Ինչ-որ) (բան) (խանգարում) (էր) (իրեն) (:)

  Richtig tokenisiert: (Ինչ-որ բան) (խանգարում) (էր) (իրեն) (:)

- **Մի քանի** ուսանողներ ներկա չէին: - ***Einige*** *Studenten waren nicht anwesend.*

  Falsch tokenisiert: (Մի) (քանի) (ուսանողներ) (ներկա) (չէին) (:)

  Richtig tokenisiert: (Մի քանի) (ուսանողներ) (ներկա) (չէին) (:)

2.2. Bewertung der automatischen Tokenisierung

Das Programm „Tokenizer“ mit den oben genannten Regeln hat bei dem Text „Corpus“ mit 121 Sätzen 100 % Genauigkeit erzielt. Mit anderen Texten kann das Ergebnis anders aussehen.

3.Automatische POS-Tagging für die armenische Sprache

Part-of-speech-Tagging (POS-Tagging) ist die Verarbeitung, in der man eine grammatikalische Rolle der Wörter in einem Satz bestimmt. Die POS-Tagging für die armenische Sprache sieht so aus.

-Beispiel-

Նրա հայրը ծրագրավորող է: —> Նրա/PP հայրը/NN ծրագրավորող/N է/VAUXSP :/P

Sein Vater Informatiker ist. PP = Personal Pronomen, NN = Nomen,

VAUX = Verb, P= Punktuationszeichen

Im Programm „Tagger“ werden simplified Tags verwendet.

Die komplette Liste ist folgendes:

1. ADP adpositions
2. ABV abbreviations
3. CC conjunction
4. CD number
5. IN preposition
6. JJ adjective
7. NN noun
8. NP noun, proper
9. PP pronoun
10. PRP participle
11. P punctuation
12. RB adverb
13. VB verb
14. VAOR aorist, verb
15. VAUX auxiliary verb
16. UH interjection
17. OP other punctuation

3.1. Training des Taggers

Bei “Natural Language Toolkit” (NLTK) verfügbare Taggers, die von einem Trainingskorpus trainiert werden, können auch beim armenischen Text verwendet werden. Im Programm „Tagger“ wird Bigram-Tagger als backoff Unigram-Tagger angewendet. Ein Beispiel der Ausgabe vom Text „Corpus“ ist:

- Beispiel der Ausgabe aus “Corpus” -

Եվ/None Հայաստանում/NP պահպանվել/None են/VAUX–*und in Armenien wurden gerettet.*

3.2. Wortformenanalyse

Für die Wörter, die nicht in den Trainningsdaten erschienen und als “None” getagt worden sind, können die Wortformenregeln einen höchstmöglichen Tag bestimmen.

Im Programm „Tagger“ wird der Regexp-Tagger von NLTK als backoff Unigram-Tagger verwendet. Die Wortarten, die von der Wortform erkannt werden können und in Programm „Tagger“ definiert sind, sind folgende:

Nomen	Suffix: -ը -ě, -ն -n
Eigenname	Präfix: Eine großgeschriebene Buchstabe
Adjektiv	Suffix: ական -akan, այինն -ajin, ավետ -avet, ավոր -avor, ավուն -avun, ե-e, յա -ya, ոտ -ot, ուն -un, ելի -eli, ալի -ali,
	Präfix: ան-an, ապ-ap, ջ-dž, տ-t, չ -č գույն -guyn, ամենա -amena
Kardinalzahl	Suffix: րորդ -rord, երորդ -erord , րդ -rd
Adverb	Suffix: բար-bar, պես-pes, որեն-oren, ովին-ovin, ակի-aki, գին-gin, պատիկ-patik, վարի-vari, ուստ-ust, ուց-uc, անց-anc, իցu-ics, ակի-aki, ացի-aci, երեն-eren
Partizip	Suffix: իս-is , ած-ac , ող – օł , իք – ik , ում – um, ու-u , ա-a , ի-i
Verb	Suffix: ել-el, ալ-al մ-m, ս-s, ի-i , նք-nk, ք-k, ն-n , ի-i, իր-ir, ր-r , ինք-ink, իք-ik, ին-in, իր-ir, ա-a, ու-u , եք-ek
	Präfix: կ - k

(die genaue Beschreibung siehe Kapitel 5)

3.3. Reihenfolgeanalyse

Im oben gezeigten Prozess können nicht erkannte Wörter durch die Wortreihenfolgeregeln höchstwahrscheinlich korrekt getagt werden, z.B. im Deutschen sind Verben immer an der zweiten Stelle.

Die Wortreihenfolge im Armenischen ist relativ flexibel, aber es lässt sich einige Regeln definieren.

Die Reihenfolgeregeln, die im Programm „Tagger“ verwendet wurden, sind folgende (genauer im Kapitel 5):

Adjektiv + Nomen

Ein Wort, das nicht im ersten und zweiten Schritt getagt worden ist, ist höchstwahrscheinlich ein **NN** (Nomen), wenn das bevorstehende Wort ein **JJ** (Adjektiv) ist.

Ավելի + Adjektiv

Ein Wort ist höchstwahrscheinlich ein Adjektiv, wenn davor das Adverb „ավելի-aveli-*mehr*“ steht.

VB + պիտի/պետք է oder պիտի/պետք է + VB

Ein Wort ist höchstwahrscheinlich ein Verb, wenn davor oder danach **„**պիտի-piti-*sollen/ müssen*“ oder „պետք է-petq ê – *sollen/müssen*“ steht.

3.4. Bewertung des automatischen POS-Taggings

Das Programm „Tagger“ kann seine Genauigkeit selbst mit einer Prozentzahl bewerten.
Mit den oben gezeigten Schritten hat „Tagger“ beim Text „Corpus“ folgendes erzielt:
Insgesammt 121 Sätze:

Trainningskorpus(Sätze)	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Prozentzahl(%)	70.7	67.8	55.7	64.7	74.0	72.0	65.1	63.0	74.0	75.0	68.8	77.4

60-65	65-70	70-75	75-80	80-85	85-90	90-95	95-100	100-105	105-110	110-115	115-120
61.9	72.1	71.4	62.1	60.0	74.3	63.0	65.3	57.3	59.6	76.0	60.6

Daher Durchschnitt **67.16%** Genauigkeit.
Für ein besseres Ergebnis wird ein größeres getagtes Textkorpus gebraucht.

4. Manuelles POS-Tagging für die armenische Sprache

Die vereinfachten Tags wurden im Kapitel 3.1 dargestellt. Dieses Kapitel enthält die gesamte Liste des POS-Tagsets mit den entsprechenden Wortarten, die für diese Arbeit erstellt wurden.

1. NN – Noun, singular
2. NNS – Noun, plural
3. NP – Noun, proper
4. JJ – adjective
5. JJS - adjective, superlative
6. CD – cardinal numbers
7. OD – ordinal numbers
8. DB – distributive numbers
9. FC – fraction numbers
10. PP – personal pronoun
11. PDEM – demonstrative pronoun
12. PREC – reciprocal pronoun
13. PINRE – interrogative and relative pronoun
14. PDE – definite pronoun
15. PIN – indefinite pronoun
16. PNG – negative pronoun
17. VB-INF – infinitiv
18. PCPS – processual participle
19. PCFR – future participle
20. PCRV – resultative participle
21. PCSB – subject participle
22. PCPT – present participle
23. PCPR – perfect participle
24. PCFRS – future participle II
25. PCNG – negative participle
26. VAOR – aorist, verb
27. SUBF – subjective, future
28. SUBP – subjective, past
29. CDF – conditional, future
30. CDP – conditional, past
31. DVF – debative, future
32. DVP – debative, past
33. IRS – imperativ, singular
34. IRP – imperativ, plural
35. VAUXSP – auxiliary verb, singular, presens
36. VAUXPP – auxiliary verb, plural, presens
37. VAUXSI – auxeliary verb, singular, imperfekt
38. VAUXPI – auxeliary verb, plural, imperfekt
39. RB – adverb
40. IN – preposition
41. CC – conjunction
42. UH – interjection
43. ADP – adpositions
44. P – punctuation
45. OP – other punctuation
46. ABV – abbreviations
47. PC – particle

5. Die einzelnen Tags mit entsprechenden Beschreibungen

5.1. Nomen

5.1.1. Nomen im Singular - NN, Nomen im Plural – NNS

In der lexikalischen Hinsicht haben Nomen als flektierte Wortart die Eigenschaften von Kasus, Numerus. Diese werden im Satz klein geschrieben. Die Nomen haben die Kategorie der Bestimmtheit und Unbestimmtheit. Die bestimmten Formen des Nomens lassen sich im Satz mit der Endung ***-ը***-ĕ/-***ն***-n erkennen: z.B. երեխա***ն,*** մարդ***ը***, կատու***ն*** (***das*** Kind, ***der*** Mensch, ***die*** Katze). Die Nomen können auch mit ***-ս***-s/-***դ***-d Endungen in bestimmter Form auftreten und entsprechend die Bedeutungen der possessiven (իմ-im-mein, քո-qo-dein), personalen (ես-es-ich, դու-du-du) und demonstrativen (այս-ays-dieses-dieser-diese) Pronomina äußern: z.B. գիրք***ս*** – ***mein*** Buch, դուք՝ հերոսներ***դ*** – ***ihr***, ***die*** Helden, ձմեռ***ս*** – ***dieser*** Winter.

Der Tag **NN** bezieht sich nicht nur auf die oben genannten Kategorien, sondern auch auf die nominalisierten Infinitive, Adjektive, Zahlen, die die Endung der bestimmten Formen bekommen und wie Nomen dekliniert werden.

Die Nomen im Plural werden mit den Endungen ***ներ***-ner, ***եր***-er, ***ք***-q, ***իկ***-ik gebildet und im Satz mit **NNS** getagt.

5.2. Eigennamen – NP

Die Eigennamen umfassen die Vornamen, Familiennamen, Ländernamen, Bergnamen usw., die im Satz immer groß geschrieben werden.

5.2.1. Adjektive – JJ

Das Adjektiv ist eine nicht-flektierte Wortart. Die Adjektive können im Satz attributive und nicht attributive Anwendungen haben. Ein Teil der Adjektive können im Satz mit ihren Affixen bestimmt werden, wie z.B. Suffixe: ***ական*** – akan, ***ային*** – ajin, ***ավետ*** – avet, ***ավոր*** – avor, ***ավուն*** – avun, ***է*** – e, ***յա*** – ya, ***ոտ*** – ot, ***ուն*** –un, ***ելի*** – eli, ***ալի*** – ali, Präfixe: ***ան*** – an, ***ապ*** – ap, ***դժ*** – dž, ***տ*** – t, ***չ*** – č.

Die komparativen Adjektive werden auch mit dem JJ getagt. Für die Bildung der komparativen Form kommt das Wort ***ավելի*** – aveli – mehr zur Hilfe, das am Anfang des Adjektivs steht. Im Satz wird der gesamte Begriff als komparatives Adjektiv verstanden, aber einzeln wird das Wort ***ավելի*** – aveli als **RB** (siehe Adverbien) und daneben stehendes

Adjektiv **JJ** getagt. Die komparativen Adjektive werden auch mit Hilfe der Konjunktion ***քան*** – kan – als gebildet. In diesem Fall wird das Wort ***քան*** – kann als **IN** (siehe Präposition) und das Adjektiv **JJ** getagt**.**

5.2.2. Adjektive, Superlativ – JJS

Die Adjektive mit dem Präfix ***ամենա*** – amena und Suffix ***գույն*** – guyn werden mit **JJS** getagt.

5.3. Zahlen

Die Zahlen, als Wortart, sind nicht-flektiert und können im Satz attributive Funktionen erfüllen.

5.3.1. Kardinalzahlen – CD

Mit **CD** werden geschriebene ganze Zahlen, ganze Zahlen in Ziffern, Jahreszahlen, Prozentzahlen getagt.

5.3.2. Ordinalzahlen – OD

Die Ordinalzahlen in der armenischen Sprache werden mit Kardinalzahlen und dem Partikel ***–րորդ*** **-** rord, ***երորդ*** **-** erord gebildet: z.B. ***երկրորդ*** **–** erkrord **–** zweite, ***հինգերորդ*** **–** hingerord **–** fünfte. Es gibt noch eine andere Schreibweise von den Ordinalzahlen, wie Kardinalzahl plus ***-րդ***-rd Partikel.

Eine Ausnahme ist die Bildung der Zahl 1. – ***առաջին*** – arajin – erste.

Den **OD** Tag bekommen auch mit römischen Buchstaben gebildete Zahlen: z.B. XX – 20., IV – 4..

5.3.3. Distributive Zahlen – DB

Es gibt zwei Bildungsweisen der distributiven Zahlen:

- mit Kardinalzahlen und Partikel ***ական***-akan: z.B. ***երկուական***-erkuakan
- mit Wiederholung zwei Kardinalzahlen: z.B. ***երկու-երկու***-erku-erku, մեկ-մեկ-mek-mek.

5.3.4. Bruchzahlen – FC

5.3.5. Pronomen

5.3.6. Personalpronomen – PP

Die Personalpronomen sind: ***ես***-es-ich, ***դու***-du-du, ***նա***-na-er, ***ինքը***-inky-er, ***մենք***-menq-wir, ***դուք***-duq-ihr, ***նրանք***-nranq-sie, ***իրենք***-irenk. Sie werden wie Nomen dekliniert, aber haben

ihre eigenen Regeln. Die deklinierten und nicht deklinierten Formen werden mit **PP** getagt. Zu den Personalpronomen gehören auch *ինքս*-inks, *ինքդ*-inkd, *ինքը*-inky, *ինքներս*-inkners, *ինքներդ*-inknerd, *ինքներր*-inknery.

5.3.7. Demonstrative Pronomen – PDEM

Die demonstrativen Pronomen können im Satz die Rolle der Substantive, Adjektive und Adverbien übernehmen. Sie können auch in der deklinierten Form auftreten und werden mit **PDEM** getagt: z.B. այս – ays – dieser, diese, dieses, նույն – nuyn – derselbe, dasselbe, dieselbe usw.

5.3.8. Reziprokpronomen – PREC

Die armenischen Reziprokpronomen sind միմյանց – mimyanc – einander, մեկմեկու – mekmeku, իրար – irar. Mit ihren deklinierten Formen werden diese Pronomen mit **PREC** Tag annotiert.

5.3.9. Interrogative und relative Pronomen – PINRE

Die interrogativen und relativen Pronomina sind die gleichen Pronomen, nur mit den anderen Funktionen im Satz. Die Interrogativen Pronomen werden für die Fragesätze benutzt, die relativen Pronomen verbinden Hauptsatz mit Nebensatz: z.B. Interrogativ - Քանի՞ մարդ էին ներկա – qani mard ein nerka – wie viel Leute waren anwesend? Նա գիտեր, թե քանի մարդ էին ներկա – na giter, te qani mard ein nerka – er wusste, wie viel Leute anwesend waren.

5.3.10. Definitives Pronomen - PDE

5.3.11. Indefinites Pronomen - PIN

5.3.12. Negative Pronomen – PNG

5.4. Verb

Das Verb, als eine Wortart in der armenischen Sprache, ist mit ihren grammatikalischen Merkmalen und Flexionsformen die reichste Wortart. Das Verb tritt im Satz mit flektierter und unflektierter Form auf. Die flektierten Formen des Verbs bekommen die grammatikalischen Merkmale von Tempus, Numerus, Modus und Personalformen. Die

unflektierten Formen dagegen haben diese Merkmale nicht. Alle Formen der Verben werden von drei Stämmen gebildet:

- Präsensstamm + entsprechende Endung
- Aorist-Stamm + entsprechende Endung
- Infinitivstamm + entsprechende Endung

5.4.1. Partizipien

Die Partizipien sind die unflektierten Formen des Verbs. Die Partizipien in der armenischen Sprache sind: Infinitiv (VB-INF), prozessual Partizip (PCPS), Partizip II resultativ (PCRV), Partizip I subjektiv (PCSB), Partizip II Futur (PCFRS), Partizip Präsens (PCPT), Partizip Perfekt (PCPR), Partizip Futur (PCFR), Partizip negativ (PGNG).

Des Weiteren werden die Formen der Partizipien ausführlicher beschrieben.

5.4.1.1. Partizipien – Infinitiv – VB-INF

Die Kategorie Infinitiv **(VB-INF)** umfasst alle Verben mit Infinitivstamm. Diese Wörter haben die Endungen ***ել***-el, ***ալ***-al: z.B. գ***ալ*** – gal – kommen, վազ***ել*** – vasel – laufen.

Die Infinitiv Verben können wie Nomen dekliniert werden (siehe Nomen).

5.4.1.2. Partizip, prozessual – PCPS

Die Prozessual Partizip Form lässt sich im Satz leicht mit der Endung իս-is erkennen, die auf den Infinitivstamm des Verbes hinzugefügt wird: z.B. երգե***լիս*** – ergelis – beim Singen, կարդա***լիս*** – kardalis – beim Lesen.

5.4.1.3. Partizip II (resultativ) – PCRV

Mit dem Tag **PCRV** können die Verben annotiert werden, bei denen dem Präsensstamm und Aorist-Stamm das Suffix ***ած***-ac hinzugefügt wird: z.B. հաջողվ***ած*** ծրագիր – hajołvac cragir – gelungenes Programm.

5.4.1.4. Partizip I (subjektiv) – PCSB

Zur Bildung dieses Partizips kommt das Suffix ***ող*** – ol zur Hilfe: z.B. կարդաց***ող*** – kardacoł – lesend, երգ***ող***-ergoł-singend.

5.4.1.5. Partizip II, im Futur – PCFRS

Diese Form des Partizips wird mit Hilfe des Infinitivstamms + des Suffixes ***իք*** – ik gebildet: z.B. կարդալ***իք*** – kardalik, խոսվել***իք*** – khosveliq.

5.4.1.6. Partizip, Präsens – PCPT

Das Präsens Partizip wird mit dem Tag **PCPT** annotiert und wird mit der Endung ***ում*** – um gebildet, die die Endungen des Infinitivs ***ել***-el, ***ալ***-al ersetzen. Diese Partizipien werden nur mit dem Hilfsverb լինել – linel – sein gebildet. (Siehe Hilfsverb im Kapitel 5.4.18.)

5.4.1.7. Partizip, Perfekt – PCPR

5.4.1.8. Partizip, im Futur – PCFR

Das Partizip Futur wird mit Hilfe des Infinitivs + der Endung ***ու***-u gebildet. Im Satz tritt es nur mit dem Hilfsverbs „sein“ auf: z.B. կարդալ***ու*** է – kardalu ê – er wird lesen, երգել***ու*** են – ergelu ên – sie werden singen.

5.4.1.9. Partizip, negativ – PCNG

Statt der Endung der Infinitiv Form ***ել***-el, ***ալ***-al werden ***ա***-a, ***ի***-i hinzugefügt. Im Satz treten die negativen Partizipien nur mit Hilfsverb (siehe die Formen des Hilfsverbs im Kapitel 5.4.18.) ***լինել*** – linel - sein in konjugierter und negativer Form auf.

5.4.2. Die flektierten Formen des Verbes

5.4.2.1. Verb, Aorist – VAOR

Diese Kategorie bezieht die Verben ein, die mit den Aorist-Stämmen + entsprechenden Personalendungen gebildet werden:

	Singular	**Plural**	**Singular**	**Plural**
1	***ի***–i–գնացի-kam	***ինք***–ink գնացիք-kamen	***ա***–a–մոռացա-vergaß	***անք***–ank–մոռացանք-vergaßen
2	***իր***–ir–գնացիր-kamst	***իք***–ik–գնացիք-kamt	***ար***–ar–մոռացար-vergaßt	***աք***–ak–մոռացաք-vergaßt
3	***Ø***-գնաց - kam	***ին***–ik–գնացին-kamen	***ավ***–av-մոռացավ-vergaß	***ան***–an–մոռացան-vergaßen

5.4.2.2. Futur, subjektiv – SUBF

Diese Kategorie der Verben wird mit dem Präsensstamm und den entsprechenden Endungen des Singulars *մ*-m, *ս*-s, *ի*-i und des Plurals *նք*-nk, *ք*-k, *ն*-n gebildet.

5.4.2.3. Past, Subjektive – SUBP

Diese Kategorie der Verben wird auch mit dem Präsensstamm und entsprechenden Endungen des Singulars *ի*-i, *իր*-ir, *ր*-r und des Plurals *ինք*-ink, *իք*-ik, *ին*-in gebildet.

5.4.2.4. Konjunktiv I (conditional Futur) - CDF

Die Verben, die mit dem Futur subjektive Form und dem Präfix կ – k gebildet werden, werden mit dem Tag **CDF** annotiert.

5.4.2.5. Konjuntkiv II (conditional, past) - CDP

Die Verben, die mit dem Past subjektive Form und dem Präfix կ – k gebildet werden, werden mit dem Tag **CDP** annotiert.

5.4.2.6. Debative, Futur - DVF

Debative Futur Form wird mit Futur subjektiv und dem Partikel պիտի-piti, պետք է-petq ê gebildet.

5.4.2.7. Debative, past - DVP

Debative Past Form wird mit Past subjektiv und dem Partikel պիտի-piti, պետք է-petq ê gebildet.

5.4.2.8. Imperativ, singular – IRS, Imperativ, plural - IRP

Der Imperativ wird in der 2. Person Singular und Plural gebildet. Der Singular wird mit dem Präsensstamm, Aorist-Stamm und den Endungen *իր*-ir, *ա*-a, *ու*-u gebildet. Die Plural Form hat nur eine Endung *եք*-ek, die immer auf Aorist-Stamm hinzugefügt wird. Bespiel: կարդա*ա* – karda – lies, գր*իր* – grir – schreib, ե կ *եք* – ekek – kommt.

5.4.3. Hilfsverb, լինել-linel-sein

- **Präsens, Singular plus Negationsformen - VAUXSP -** եմ-em-bin, ես-es-bist, է-ê-ist, չեմ-čem-bin nicht, չես-čes-bist nicht, չէ-čê-ist nicht.

- **Präsens, Plural plus Negationsformen - VAUXPP –** ենք-enk-sind, եք-ek-seid, են-en-sind, չենք-čenk-sind nicht, չեք-ček-seid nicht, չեն-čen-sind nicht.
- **Imperfekt, Singular plus Negationsformen -VAUXSI –** էի-êi-war, էիր-êir-warst, էր-êr-war, չէի-čêi-war nicht, չէիր- čêir-warst nicht, չէր- čêr-war nicht.
- **Imperfekt, Plural plus Negationsformen - VAUXPI –** էինք-êink-waren, էիք-êiq-wart, էին-êin-waren, չէինք-čeink-waren nicht, չէիք-čeik-wart nicht, չէին-čein-waren nicht.

5.5. Adverb – RB

Das Adverb ist eine nicht flektierte Wortart. Die Adverbien sind im Satz entweder einfache oder von Substantiven und Adjektiven durch Suffixe abgeleitete Wörter: բար-bar, պես-pes, որեն-oren, ովին-ovin, ակի-aki, գին-gin, պատիկ-patik, վարի-vari, ուստ-ust, ուց-uc, անց-anc, իցu-ics, ակի-aki, ացի-aci, երեն-eren usw. Die werden im Satz mit **RB** getagt.

5.6. Präposition – IN

Die Präpositionen sind meist nicht flektierbar und werden im Satz mit dem Tag **IN** getagt. Die Präpositionen in der armenischen Sprache haben fast die gleiche Rolle, wie die Kasusendungen und oft können die die letzteren ersetzen. Sie können dem Wort entweder vor oder nach gestellt werden: z.B. սեղանի ***վրա*** – sełani vra – auf dem Tisch, սեղան***ին*** – sełanin – auf dem Tisch.

5.7. Konjunktionen – CC

Die Konjunktionen gehören zu den nicht flektierbaren Wortarten. Sie verbinden Sätze und Wörter zusammen: z.B. ու-u-und, իսկ-isk-aber, որպեսզի-orpeszi-damit.

5.8. Interjektion – UH

Die Interjektionen sind nicht flektierbar und äußern im Satz die Gefühle der Sprechenden. Die treten im Satz meist mit einem armenischen Satzzeichen (՝, siehe Kapitel 2.1.) auf.

5.9. Adposition - ADP

5.10. Punktuation – P (siehe Kapitel 2.1.)

5.11. Andere Punktuationen – OP (siehe Kapitel 2.1.)

6. Die Disambiguierung der Wörter

In der armenischen Sprache ist die Zahl der mehrdeutigen Wörter ziemlich hoch. Solche Wörter haben die gleiche Schreibweise und man kann nur im Kontext bestimmen, zu welcher Wortart das Wort gehört. Des Weiteren wird das Phänomen anhand einiger Bespiele näher beschrieben.

- **NN oder JJ**

Viele Wörter, die die gleiche Schreibweise haben, können im Satz entweder als Nomen oder als Adjektive angewendet werden.

Beispiel: **Հիվանդ/JJ** տղամարդը… – **hivand** tłamardě-der **kranke** Mann

Մի **հիվանդ/NN…** – **mi hivand-ein Kranke**

Երիտասարդ/JJ ուժեր… – eritasart užer-**junge** Kräfte

Երկու **երիտասարդ/NN…** – erku **eritasard** – zwei **Jungen**

- **IN oder NN**

Wörter, die als Präpositionen oder Nomen auftreten:

Beispiel: Հանդիպման **ժամանակ/IN…** – handipman **žamanak-während** des Treffs

Մի **ժամանակ/NN** առաջ… – mi **žamanak** araj – vor einer **Zeit**

Աղջկա **երեսից/JJ** չկարողացանք... – ałǰka **eresic** č'karołacank - **wegen** des Mädchens

Աղջկա **երեսից/NN** երևում էր ... - ałǰka **eresic** erevum er – Vom **Gesicht*** des Mädchens sah man…

- **IN oder RB**

Wörter, die als Präpositionen oder Adverbien benutzt werden können.

Beispiel: Դարեր **առաջ/IN** – darer **araj** – **vor** Jahren

Գնալ **առաջ/RB** – gnal **araj - vor**gehen

Քեզանից **հետո/IN** – qezanic **heto** – **nach** dir

Հետո/RB կգամ – **heto** kgam – ich komme **nachher**

- **JJ oder RB**

Die Adjektive und Adverbien können sich voneinander dadurch unterscheiden, dass Adjektive attributive Funktionen erfüllen und meist vor Nomen stehen und die Adverbien prädikative.

Beispiel: **Շատ/RB** խոսել – **šat** xosel – **viel** sprechen

Շատ/JJ գրքեր **- šat** grker - **viele** Bücher

- **IN oder PCRV**

Das resultative Partizip kann attributive Funktionen erfüllen und vor einem Nomen stehen. Das gleiche Wort kann auch die Rolle einer Präposition einnehmen:

Beispiel: **Սկսած**/PCRV գործ – **sksac** gorc – **angefangene** Arbeit

Այդ օրից **սկսած/IN** – ayd oric **sksac – ab** diesem Tag

- **CD oder PNG**

Es geht hier um das Wort մի – mi – ein/eine, dass als eine Kardinal Zahl oder mit einem anderen Wort zusammen als ein unbestimmtes oder ein negatives Pronomen verwendet wird.

Beispiel: **Մի**/CD աշակերտ կար միայն–mi ašakert kar miayn-es gab nur einen Schüler

Ոչ մի/PNG մարդ - oč’ mi mard – kein Mensch

7. Zusammenfassung

Das POS-Tagging bietet viele Einsatzmöglichkeiten bei der maschinellen Textverarbeitung. Das Ziel dieser Arbeit war es automatische und manuelle Guidelines der POS-Tagging für die armenische Sprache zu erstellen. Dafür wurde eine Liste von POS-Tagsets für die manuelle und eine Liste von simplified POS-Tagsets für die automatische Wortartenerkennung erstellt. Für die weitere Entwicklung dieses Bereichs wird vor allem ein größerer Textkorpus nötig sein, der ermöglicht, mehr Flexionsformen der Morphologie zu entdecken. Dabei können auch neue Reihenfolgeregeln für die automatische Wortartenerkennung hilfreich sein. Darüber hinaus müssen die Kategorien der Texte eine gewisse Vielfalt haben und mit dem Training von solchen größeren Textmengen kann man die Genauigkeit des Taggers erhöhen.

In der Arbeit wurden auch die wichtigsten Guidelines der Segmentierung besprochen, die auch bei der Vielfalt der Textkategorien verbessert werden können. Sie können ihren Beitrag für die Erhöhung der Genauigkeit des POS-Taggings leisten.

Ein Bereich für den Einsatz des POS-Taggings könnte die maschinelle Übersetzung sein, wo das Programm nicht nur die grammatikalische Rolle des Wortes erkennt, sondern auch eine passende Übersetzung von mehreren Bedeutungen des Wortes durch den Kontext des Satzes auswählt.

Literaturverzeichnis

- Beatrice Santorini: Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision), Pennsylvania, 1990
- Anne Schiller, Simone Teufel, Christine Stöckert: Guidelines für das Tagging deutscher Textcorpora mit STTS, Stuttgart, 1999
- Ezekyan: Հայոց լեզու – Armenische Sprache, Jerewan, 2007
- Jasmine Dum-Tragut: Modern Eastern Armenian, 2009
- Wikipedia:
  http://hy.wikipedia.org/wiki/%D5%80%D5%A1%D5%B5%D5%A1%D5%BD%D5%BF%D5%A1%D5%B6
  http://hy.wikipedia.org/wiki/%D5%83%D5%A1%D5%BA%D5%B8%D5%B6%D5%AB%D5%A1