Universität Zürich[UZH]

Institut für Computerlinguistik

Betreuer: Simon Clematide

Morphologieanalyse für Rumantsch Grischun

Praktikum intern und Qualifikationsarbeit ohne Veranstaltung

Reto Baumgartner

28. Juli 2013

Art des Projekts

Bestandteile:

- Übungsleitung in «Finite-State-Methoden in der Sprachtechnologie» (FS13)
- Programmierprojekt
- Zeit pro ECTS als Kriterium

Kurzportrait: Rätoromanisch

5 traditionelle Schriftidiome und Rumantsch Grischun

Aus «Der Rabe und der Fuchs» von Jean de La Fontaine [3, S. 31]:

Sursilvan: Cheu ha ella viu sin in pegn in tgaper che teneva in toc caschiel en siu bec.

Sutsilvan: Qua â ella vieu sen egn pegn egn corv ca taneva egn toc caschiel ainten sieus pecel.

Surmiran: Cò ò ella via sen en pegn en corv tgi tigniva en toc caschiel ainten sies pechel.

Puter: Cò ho'la vis sün ün pin ün corv chi tgnaiva ün töch chaschöl in sieu pical.

Vallader: Qua ha'la vis sün ün pin ün corv chi tgnaiva ün toc chaschöl in seis pical.

RG: Qua ha ella vis sin in pign in corv che tegneva in toc chaschiel en ses pichel.

Unterschiede zwischen den Varietäten I

Analytische Zeitformen der Verben:

Rumantsch Grischun	Puter [4, S. 9–10]
Präsens Indikativ	Präsens Indikativ
Imperfekt Indikativ	Imperfekt Indikativ
	Einfaches Perfekt Indikativ
	Futur Indikativ
	Futur Dubitativ
Konjunktiv	Konjunktiv I
Konditional	Konjunktiv II

Unterschiede zwischen den Varietäten II

Weitere wichtige Unterschiede:

- Verneinung
- Pronomina
- Zusammenzug aus Präposition und Artikel

Designentscheide

Tags

Empfehlungen von Beesley und Karttunen [1]:

- Nichts aufzwingen – verwenden, was schon üblich ist
- Konsistenz: Gleiche Tags für gleiches Verhalten
- Verwende die Tags, die Xerox verwendet
- Entscheide dich vorher für Tags und ihre Reihenfolge

Folgen für dieses Projekt I

Reihenfolge der Tags:

- wie Französisch:
  «pled+Masc+SG+Noun»
- wie Italienisch, Spanisch und Portugiesisch:
  «pled+Noun+Masc+Sg»

Italienisch ist verbreitet in Graubünden

Folgen für dieses Projekt II

Neuer Tag «+AccDat»

- Keine reine Dativpronomen in RG
- Daneben reine Akkusativformen
- Analogie zum Tag «+MF»

Folgen für dieses Projekt III

Begriffe in der Grammatik von Caduff et al. [2]:

- Konjunktiv: «+Conj» statt «+PresSubj»
- Personalpronomen: «+Pron+Pers» oder «+Pron»
- Indefinitpronomen: «+Pron+Indef»

Besonderheiten

Welches Lemma?

- Pronomina: «jau», «ti», «sai»
- Pluraliatantum: mit Pluralendung

Genusmarkierung bei Personalpronomina:

- +Masc
- +Fem
- +MF

Verben mit suffigierten Pronomina:
gidar+Verb+PresInd+3P+Sgˆ|+Pron+Pers+Nom+3P+Masc+Sg
gida'l

Tokenisierung

Mehrworttoken:

- Eher strenge Trennung wie bei Xerox

Apostrophierung:

- Apostroph beim ersten Teil → Trennung
  l'onn → l' + onn
- Apostroph beim zweiten Teil → keine Trennung
  gida'l → gida'l

Problem: Analyse der Idiome

Lösung Ersetzungsregeln

Typische lautliche und schriftliche Unterschiede zwischen den Idiomen und RG

Lemmatisierung in RG

Probleme:

- Verschiedene Grammatische Kategorien
- Verschiedenes Vokabular
- Gefahr der Übergenerierung
- Gefahr der Überlastung

Lösung Überschreibung

Beschränkter Einsatzbereich

Hier für die abweichenden Formen der:

- Artikel
- Pronomina
- Präpositionen (bei falschen Freunden)

Tests

Testverfahren

Empfehlungen von Beesley und Karttunen [1]:

- Auf realen Korpora
- Überprüfen des Alphabets
- Testen mit Minus

Testkorpora

Rumantsch Grischun

- Sauberes Korpus: Allegra-Korpus [1]
- Unsauberes Korpus: Empfohlene Artikel der rätoromanischen Wikipedia [2]

Schriftidiome:

- Unsaubere Korpora: Nach Schriftidiom geordnete Artikel der rätoromanischen Wikipedia

[1] Erhältlich über `http://www.latl.unige.ch/allegra/` (letzter Zugriff 2013-07-25)

[2] Zu finden unter `http://rm.wikipedia.org/` (letzter Zugriff 2013-07-25)

Ergebnisse für Rumantsch Grischun

Korpus	Korpusgrösse	Unbekannte	Recall
Allegra	1 312 857	ohne	92.62%
Allegra	1 312 857	mit	97.53%
Wikipedia	450 727	ohne	83.31%
Wikipedia	450 727	mit	94.76%

Ergebnisse für die Schriftidiome

Idiom	Korpusgrösse	Recall
Puter	15 678	68.70%
Surmiran	4 524	70.36%
Sursilvan	23 056	72.70%
Sutsilvan	1 039	66.41%
Vallader	6 771	72.47%

Weiterführende Ideen

Weiterführende Ideen

- Separate Integration der Schriftidiome
- Ergänzung oder Ersetzung der Wortlisten
- Ergänzung sprachunabhängiger Teile
- Verarbeitung von Namen

Literatur

Literatur

[1] Kenneth R. Beesley und Lauri Karttunen. *Finite-State Morphology: Xerox Tools and Techniques*. The Document Company—Xerox, 2000.

[2] Renzo Caduff, Uorschla N. Caprez und Georges Darms. *Grammatica d'instrucziun dal rumantsch grischun*. Dissertation, Seminari da rumantsch da l'Universitad da Friburg, Fribourg, 2006.

[3] Lia Rumantscha (Hrsg.). Facts figures. aus dem deutschen von daniel telli. 2., überarbeitete und aktualisierte ausgabe. 2004.

[4] Gion Tscharner. (kein titel). URL `http://www.udg.ch/dicziunari/files/grammatica_puter.pdf` (letzter Zugriff: 2013-07-24). Grammatik für Puter.

Fragen