# Morphologieanalyse für Rumantsch Grischun

Reto Baumgartner, Martina Bachmann, Rolf Badat, Daniel Hegglin,
Susanna Tron, Melanie Widmer
Universität Zürich
Institut für Computerlinguistik
26. Juli 2013

## 1 Abstract

Als Gruppenarbeit wurde an der Universität ein Morphologieanalysesystem für die schweizerische Landessprache Rätoromanisch erstellt. Dafür wurde die Standardvarietät Rumantsch Grischun gewählt und mit Hilfe von Finite-State-Methoden implementiert. Auch die traditionellen Standardvarietäten des Rätoromanischen lassen sich bis zu einem gewissen Grad mit dem gebauten System behandeln. Die linguistischen Teile orientieren sich eng an existierenden Systemen nahe verwandter Sprachen.

## 2 Ausgangslage

Als Grundlage für die Wortbildung diente die Grammatik von Caduff et al. [2].

Als Basis für die Wortlisten wurde das Pledari grond online der Lia Rumantscha [4] verwendet, wobei die Verben und Substantive systematisch gesammelt werden konnten.

Die Wahl der Tags folgte den Empfehlungen von Beesley und Karttunen [1, S. 335–366]. Für Zweifelsfälle wurde auch das Online-Morphologieanalysesystem von Corporation [3] für die italienische Sprache hinzugezogen.

## 3 Installation

Das Morphologieanalysesystem für Rumantsch Grischun lässt sich bequem mit Hilfe vom Makefiles installieren. Die Voraussetzung für die Installation sind die Finite-State-Werkzeuge von Xerox (`xfst` und `lexc`)[1] oder die Open-Source-Variante Foma (mit `foma` und `lexc`)[2].

1. Erhältlich über `http://www.stanford.edu/~laurik/fsmbook/home.html` (letzter Zugriff: 2013-07-24)
2. Erhältlich über `http://code.google.com/p/foma/` (letzter Zugriff: 2013-07-24)

Für die traditionellen Schriftidiome existiert ein Behelf, der mit ein paar wenigen gelisteten Formen und regelmässigen Ersetzungen von Buchstaben oder Buchstabengruppen im Rumantsch Grischun die Formen der traditionellen Schriftidiome bildet. So können aber natürlich nicht alle Formen erkannt werden, nur schon dadurch, dass die Schriftidiome sich in der Grammatik manchmal deutlich vom Rumantsch Grischun unterscheiden.

Für die Installation müssen die Dateien des Archivs im gewünschten Ordner entpackt werden und dort können mit folgenden Kommandos die Netzwerke wiederverwendbar gespeichert werden:

| | |
|---|---|
| `make` | (für die Installation mit `xfst`) |
| `make -f Makefile-foma` | (für die Installation mit `foma`) |
| `make -f Makefile-idioms` | (Erkennung der Schriftidiome, mit `xfst`) |
| `make -f Makefile-idioms-foma` | (Erkennung der Schriftidiome, mit `foma`) |

Mit diesen Kommandos können die Netzwerke nach Änderungen in den Wörterlisten oder der weiteren Verarbeitung aktualisiert werden.

## 4 Benutzung

Die bei der Installation erstellten Dateien mit `.fst` können in `xfst`/`foma` geladen werden und dort weiterverwendet werden mit:

```
xfst[0]: load stack GrischunGuessing.fst
```

oder sie können auf der Kommandozeile für die Analyse mittels `lookup`/`flookup` verwendet werden:

```
$ lookup Grischun.fst < tokenis-Infile.txt > Outfile.txt
```

## 5 Verwendete Tags

### 5.1 Wortartentags

| | |
|---|---|
| +Adj | Adjektiv |
| +Adv | Adverb |
| +Art | Artikel |
| +Conj | Konjunktion |
| +Dig | Zahlen in Ziffernschreibung |
| +Initial | Initialenabkürzungen wie A. |
| +Interj | Interjektion |
| +Let | Buchstabe |
| +Noun | Substantiv |
| +Num | Zahlwörter |
| +Prep | Präposition |
| +Pron | Pronomen |
| +Prop | Namen |
| +Punc | Satzzeichen |
| +PUNCT | weitere Zeichen im Satz |
| +Rom | römische Zahlen |
| +Subj | Subjunktionen |
| +Verb | Verb |

### 5.2 Genauere Einteilung der Wortarten

Pronomen:

| | |
|---|---|
| +Pron +Dem | Demonstrativpronomen |
| +Pron +Indef | Indefinitpronomen |
| +Pron +Interrog | Interrogativpronomen |
| +Pron +Pers | Personalpronomen |
| +Pron +Poss | Possessivpronomen |
| +Pron +Refl | Reflexivpronomen |

Zahlen:

| | |
|---|---|
| +Dig +Card | Kardinalzahlen in Ziffern |
| +Dig +Dec | Dezimalzahlen |
| +Dig +Degree | Gradangaben |
| +Dig +Ord | Ordinalzahlen in Ziffern |
| +Dig +Percent | Prozentzahlen |
| +Num +Card | Kardinalzahlen |
| +Num +Ord | Ordinalzahlen |
| +Num +Adj | Multiplikativzahlen |
| +Rom +Card | römische Kardinalzahlen |
| +Rom +Ord | römische Ordinalzahlen |

Satzzeichen:

| | |
|---|---|
| +Punc +Beg | öffnende Satzzeichen |
| +Punc +Mid | mittlere Satzzeichen |
| +Punc +End | schliessende Satzzeichen |

Abkürzungen:

| | |
|---|---|
| +Noun +Abbr | Abkürzungen von Substantiven |

### 5.3 Deklination und Konjugation

Kasus:

| | |
|---|---|
| +Nom | Nominativ |
| +Acc | Akkusativ |
| +AccDat | Akkusativ oder Dativ |

Numerus:

| | |
|---|---|
| +Sg | Singular |
| +Pl | Plural |

Genus:

| | |
|---|---|
| +Fem | feminin |
| +Masc | maskulin |
| +MF | maskulin oder feminin |

Person:

| | |
|---|---|
| +1P | erste Person |
| +2P | zweite Person |
| +3P | dritte Person |

Definitheit:

| | |
|---|---|
| +Def | bestimmt |
| +Indef | unbestimmt |

Steigerung:

| | |
|---|---|
| +Comp | unregelmässiger Komparativ |
| +Sup | absoluter Superlativ |

Betontheit:

| | |
|---|---|
| +Aton | unbetont |
| +Ton | betont |

Verbformen:

| | |
|---|---|
| +PresInd | Präsens Indikativ |
| +ImpInd | Imperfekt Indikativ |
| +Conj | Konjunktiv |
| +Cond | Konditional |
| +Impv | Imperativ |
| +Inf | Infinitiv |
| +Gerund | Gerundium |
| +PastPart | Partizip Vergangenheit |

### 5.4 Weitere Tags

Komposition:

| | |
|---|---|
| ^DB | Derivationsgrenze |
| ^\| | Grenze vor Suffigierung |
| ^= | Kompositionsgrenze (ausser Substantive) |

Diverse:

| | |
|---|---|
| * | Grossschreibung |
| +UNKNOWN | Unbekannte Form |
| +Apo | Apostrophierte Form oder mit Hiatustilger |

Tie Tags +UNKNOWN und * können in `collection-RG.xfst` geändert werden. Für der Kompilierung mit den Schriftidiomen können die Tags am Beginn der Datei `collection.xfst` geändert werden.

## 6 Wortarten

## 6.1 Adjektive

Adjektive werden wie folgend markiert:

| Lemma | Wortart | Steigerungsstufe | Genus | Numerus |
|---|---|---|---|---|
| bun | +Adj | | +Masc | +Sg |
| | | +Comp | +Fem | +Pl |
| | | +Sup | | |

Die Markierung für den Komparativ wird nur für die unregelmässige Steigerung verwendet. Gleichzeitig steht er auch, wenn eine entsprechende Adjektivform superlativisch verwendet wird. Die Markierung für den Superlativ steht für Formen mit der Endung ‹-ischem› die nicht eine Steigerungsform im engen Sinn, sondern eine Intensivierungen des Adjektivs beinhaltet. Für den Positiv steht keine Markierung.

Die Integration der Adjektive findet in `adj/adj.xfst` statt. Es wird eine Aufteilung der Adjektive in verschiedene Kategorien verwendet.

### 6.1.1 Regelmässige Adjektive

Wie regelmässige Adjektive (wie *calm* – *calma*) werden auch die Adjektive mit Konsonantenverdoppelung vor der femininen Endung (wie *brut* – *brutta*) und Adjektive mit flüchtigem Vokal (wie *liber* – *libra*) behandelt. Durch eine vorausgehende Behandlung können alle schliesslich wie regelmässige Adjektive behandelt werden. Die drei Adjektivuntergruppen sind einzeln in folgenden Dateien gelistet:

- `wordlists/adj-reg.txt` für die ganz regelmässigen Adjektive. Diese Liste ist noch weit entfernt von der Vollständigkeit.
- `wordlists/asj-doubling.txt` für die Adjektive mit Konsonantenverdopplung. Auch diese Liste sollte noch erweitert werden.
- `wordlists/adj-e.txt` für die Adjektive mit flüchtigem Vokal.

### 6.1.2 Adjektive mit Partizipendung

Diese Adjektive enden in -à oder -ì (z. B. *affectuà* – *affectuada* oder *partì* – *partida*). Die meisten von ihnen sind auch Partizipien, jedoch solche, die im Pledari Grond als Lemma gelistet sind.

Diese Adjektive sind gelistet in:

- `wordlists/adj-part.txt`. Die Liste kann als ziemlich vollständig angesehen werden.

### 6.1.3 Unveränderliche Adjektive

Für die unveränderlichen Adjektive wurden die gleichen Tags verwendet wie für die regelmässigen Adjektive. Somit ist die Analyse nie eindeutig möglich, aber die Einheitlichkeit ist bewahrt. Auf den Superlativ wurde verzichtet, da nicht klar ist, ob und wie dieser gebildet werden könnte. Die unveränderlichen Adjektive sind gelistet in der Datei:

- `wordlists/adj-inv.txt`

### 6.1.4 Unregelmässige Adjektive

Die unregelmässigen Adjektive teilen sich in zwei Gruppen auf, nämlich in diejenigen mit einer unregelmässigen Steigerung und diejenigen mit einer unregelmässigen Formenbildung. Die Formen sind komplett in `lexc` geschrieben und überschreiben die anderen Formen, wenn sie die gleiche Oberseite aufweisen. Nebeneinanderstehende Formen sollten deshalb alle gelistet werden. Die Implementierung findet in folgenden Dateien statt:

- `adj/adj-irr.lexc` für die unregelmässige Formenbildung.
- `adj/adj-comp-irr.lexc` für die unregelmässige Steigerung.

### 6.1.5 Hypothetische Formen

Die Erratung von unbekannten Adjektivformen ist nur bei den regelmässigen Adjektiven (inkl. Konsonantenverdoppelung und flüchtigen Vokal) gemacht. Die Adjektive mit Partizipendung wurden bewusst weggelassen, da solche Formen in erster Linie eher Verbformen sind und so einerseits schon integriert sind, andererseits auch bereits in den meisten Fällen korrekt analysiert werden können.

## 6.2 Adverbien

Für die Adverbien dienen folgende Markierungen:

| Lemma | Wortart | Steigerungsstufe | Derivationsgrenze | Wortart |
|---|---|---|---|---|
| bun | +Adj | | ^DB | +Adv |
| | | +Sup | | |
| main | +Adv | | | |

Die oben gelistete Behandlung wie bei *bun* behandelt Adverbien, die von Adjektiven abgeleitet sind. Die untere Art zeigt, wie Kurzadverbien behandelt werden. Die Adverbformen werden in `adv/adv.xfst` gesammelt. Die Implementierung der Formen geschieht analog zu den regelmässigen Adjektiven (Kapitel 6.1.1) und den Adjektiven mit Partizipendung (Kapitel 6.1.2). Auf die Behandlung der unregelmässigen Formen und der unveränderlichen muss hier aber weiter eingegangen werden.

### 6.2.1 Adverbien aus unveränderlichen Adjektiven

Diese Lemmata sind in folgender Liste gesammelt:

- `wordlists/adv-adj.txt`. Die Liste muss möglicherweise erweitert werden.

### 6.2.2 Unregelmässige Adverbien

Adjektive, welche die feminine Form unregelmässig bilden, zeigen dieses Verhalten auch bei den Adverbien (z. B. lartg – largia – larigamain). Diese Formen sind komplett in `lexc` geschrieben und überschreiben regelmässige Formen, die die gleiche Oberseite aufweisen:

- `adv/adv-irr.lexc`

### 6.3 Artikel

Die Artikel und Präpositionalartikel werden mit folgenden Tags genauer bezeichnet:

| lemma | Wortart | Grenze | Wortart | Bestimmth. | Genus | Numerus | Endung |
|---|---|---|---|---|---|---|---|
| in | +Art | | | +Def | +Masc | +Sg | |
| | | | | +Indef | +Fem | +Pl | +Apo |
| da | +Prep | ^= | +Art | | | | |

Diese Formen sind komplett in `lexc` gelistet und in der Datei `art-pron/art.lexc` zu finden. Hier ist keine Erweiterung nötig oder vorgesehen.

### 6.4 Buchstaben und Initialen

Als Initialen zählt die Kombination aus einem Grossbuchstaben mit einem Punkt. Sie werden mit +Initial gekennzeichnet. Buchstaben sind dagegen Minuskel und Majuskel und sie werden mit +Let gekennzeichnet. Als Kriterium für die Wahl der Buchstaben wurden die Zeichensätze ISO 8859-1 und ISO 8859-15 gewählt und die Buchstaben daraus kombiniert.

Die Buchstaben und Initialen sind in `particles/letter.lexc` gelistet.

### 6.5 Interjektionen

Die Interjektionen tragen den Tag +Interj und sie sind in `particles/interj.lexc` gelistet.

### 6.6 Interpunktion

Für die Interpunktion dienen folgende Tags:

| Lemma | Wortart | Unterart |
|---|---|---|
| , | +Punc | |
| | | +Beg |
| | | +Mid |
| | | +End |
| % | +PUNCT | |

Satzzeichen und weitere Interpunktionszeichen sind in `particles/interpunct.lexc` gelistet. Satzzeichen tragen den Tag +Punc und, falls es sich um öffnende oder schliessende Zeichen handelt, den Tag +Beg oder +End. Die dritte Unterteilung (+Mid) steht, wenn das Zeichen für gewöhnlich zwischen zwei Einheiten steht, die es verbindet.

Der Tag +PUNCT steht bei Zeichen, die grundsätzlich nicht für die Strukturierung eines Satzes verwendet werden, aber dennoch sehr häufig auftreten.

### 6.7 Konjunktionen und Subjunktionen

Es wird unterschieden zwischen Konjunktionen (+Conj) und Subjunktionen (+Subj). Apostrophierte Formen oder solche mit Hiatustilger tragen zusätzlich den Tag +Apo . Die Konjunktionen und Subjunktionen sind in `particles/conj.lexc` gelistet.

### 6.8 Numerale und Zahlen

Für Zahlen und Zahlwörter stehen folgende Tags:

| Lemma | Zahlart | | Mass | Genus | Numerus |
|---|---|---|---|---|---|
| 123 | +Dig | +Card | | | |
| | | | +Percent | | |
| | | | +Degree | | |
| 124 | +Dig | +Ord | | | |
| | | | | +Masc | +Sg |
| | | | | +Fem | +Pl |
| 1.67 | +Dig | +Dec | | | |
| in | +Num | +Card | | +MF | |
| | | | | +Masc | |
| | | | | +Fem | |
| sis | +Num | +Ord | | +Masc | +Sg |
| | | | | +Fem | +Pl |
| in | +Num | +Adj | | +Masc | +Sg |
| | | | | +Fem | +Pl |
| II | +Rom | +Card | | | |
| II | +Rom | +Ord | | | |
| | | | | +Masc | +Sg |
| | | | | +Fem | +Pl |

Die Numerale und Zahlen sind in `num/num.xfst` implementiert.

Die Ordnungszahlen tragen Tags für die Deklinationen, wenn sie mit dem Ordinalzahlensuffix «-avel» gebildet werden. Werden sie hingegen mit Punkt gebildet, dann können keine Deklinationsangaben gemacht werden.

Bei den Netzwerken wird unterschieden zwischen Zahlen und Zahlwörtern. Während die Zahlen allgemein gültig sind, sind Zahlwörtern schriftidiombedingten Wechseln unterworfen.

### 6.9 Präpositionen

Präpositionen werden mit dem Tag +Prep markiert. Bei Apostrophierung oder Hiatustilger steht zusätzlich der Tag +Apo . Die Präpositionen sind in `particles/prep.lexc` gelistet.

Zur Kombination aus Artikel und Präposition steht mehr bei 6.3.

### 6.10 Pronomina

Die morphologischen Angaben zu Pronomina werden durch folgende Tags gegeben:

| Lemma | Wortart | Unterart | Kasus, Ton | Pers. | Genus | Num. | Endung |
|---|---|---|---|---|---|---|---|
| jau | +Pron | +Pers | +Nom | +1P | +Masc | +Sg | |
| sai | | +Refl | +Acc +Ton | +2P | +Fem | +Pl | +Apo |
| | | | +AccDat +Aton | +3p | +MF | | |
| mes | +Pron | +Poss | | | +Masc | +Sg | |
| | | | | | +Fem | +Pl | |
| lez | +Pron | +Dem | | | | | |
| tgi | | +Interrog | | | +Masc | +Sg | +Apo |
| tut | | +Indef | | | +Fem | +Pl | |

Bei den Demonstrativ-, Interrogativ- und Indefinitpronomina stehen Deklinationsendungen nur bei veränderlichen Lemmata. Die Possessivpronomina können zu Substantiven deriviert werden. Dabei steht der Tag ^DB und die restlichen Tags wie bei den Substantiven.

Die Pronomina sind in `art-pron/pron.lexc` gelistet.

### 6.11 Substantive

Die Substantive werden durch folgende Tags bestimmt:

| Lemma | Wortart | Genus | Numerus |
|---|---|---|---|
| pled | +Noun | +Masc | +Sg |
| | | +Fem | +Pl |

Die Integration der Substantive findet in `noun/noun.xfst` statt. Die Substantive sind in folgende Gruppen eingeteilt: Regelmässige Substantive je nach Genus, Pluraliatantum und Singulariatantum je nach Genus, maskuline Substantive auf die Partizipendungen -à und-ì, sowie auf die Endung -è. Die mit Bindestrich zusammengesetzten Komposita werden hier mitbehandelt, die Komposita ohne Bindestrich weichen in der Deklination nicht ab. Die unregelmässigen Substantive sind separat in `lexc` integriert.

#### 6.11.1 Regelmässige Substantive

Die regelmässigen Substantive sind in folgenden Dateien abgelegt:

- `wordlists/noun-fem.txt` für die femininen Substantive.
- `wordlists/noun-masc.txt` für die maskulinen Substantive. In diesen beiden Listen könnten noch Singulariatantum enthalten sein. Dies hat aber nur Folgen, wenn das Analysetool als Akzeptor verwendet werden soll, da in den anderen Fällen der Input eine ausreichende Beschränkung darstellt.

#### 6.11.2 Singulariatantum und Pluraliatantum

Als Singulariatantum wurden die Wörter von Caduff et al. [2] übernommen und ergänzt. Als Pluraliatantum dienen die Formen, die im Pledari grond als Lemmata im Plural vorkommen. Aus diesem Grund erscheint auch hier der Plural im Lemma. Die Singulariatantum und Pluraliatantum sind in folgenden Listen gesammelt:

- `wordlists/noun-fem-sing.txt` für die femininen Singulariatantum.
- `wordlists/noun-masc-sing.txt` für die maskulinen Singulariatantum.
- `wordlists/noun-fem-plur.txt` für die femininen Pluraliatantum.
- `wordlists/noun-masc-plur.txt` für die maskulinen Pluraliatantum.

#### 6.11.3 Substantive auf -à, -ì und -è

Diese maskulinen Substantive ändern ihre Endung, bevor die Endung für den Plural hinzukommt (mantè – mantels, marì – marids). Sie stehen in einer Liste, da sie sich problemlos gemeinsam behandeln lassen:

- `wordlists/noun-part.txt`

#### 6.11.4 Unregelmässige Substantive

Die unregelmässigen Substantive sind in `lexc` geschrieben und überschreiben Formen mit derselben Oberseite. Sie liegen in der Datei:

- `noun/noun-irr.txt`

### 6.11.5 Hypothetische Formen

Die Verarbeitung für unbekannte Formen enthält die regelmässigen Substantive, die Substantive auf -è und die Komposita mit diesen Formen. Von den Substantiven mit Paritzipendung wurde abgesehen, da diese schon bei den Verben integriert sind, sodass eine brauchbare Analyse möglich ist.

### 6.11.6 Abkürzungen und Namen

In `wordlists/noun-abbr.txt` sind Abkürzungen für Substantive enthalten. Sie tragen die Tags +Noun +Abbr . Ist eine Abkürzungsliste vorhanden, empfiehlt es sich, diesen Teil zu ersetzen.

In `wordlists/noun-proper.txt` sind Namen gelistet. Für Personennamen liegt es nahe, aus bestehenden System diesen Teil zu übernehmen. Für sprachspezifische Namen werden aber spezifische Listen vonnöten sein.

## 6.12 Verben

Die morphologischen Angaben zu den Verben werden mit folgenden Tags dargestellt:

| Lemma | Wortart | Form | Person | Genus | Numerus |
|---|---|---|---|---|---|
| midar | +Verb | +PresInd | +1P | | +Sg |
| | | +ImpInd | +2P | | +Pl |
| | | +Cond | +3P | | |
| | | +Conj | | | |
| | | +Impv | | | |
| midar | +Verb | +Inf | | | |
| | | +Gerund | | | |
| midar | +Verb | +PastPart | | +Masc | +Sg |
| | | | | +Fem | +Pl |

Zusätzlich können noch Endungen folgen, wenn das Verb von Pronomina gefolgt ist. Folgt das Pronomen *ins* wird entweder die Verbendung apostrophiert oder ein ‹n› suffigert, was beides mit +Apo markiert wird.

Die Personalpronomina werden hingegen direkt an das Verb suffigiert und die Verbindungsgrenze mit ^| markiert. Danach folgen die üblichen Angaben der Pronomina:

gidar+Verb+PresInd+3P+Sg^|+Pron+Pers+Nom+3P+Masc+Sg gida'l

Die Implementierung der Verben erfolgt in `verb/verb.xfst` und es wird nach drei Verbgruppen unterschieden: Regelmässige Verben, Verben mit Vokalwechsel und unregelmässige Verben. Die Bildung der unregelmässigen Partizipformen erfolgt separat, da diese nicht dem gleichen Aufteilungsschema folgen.

### 6.12.1 Regelmässige Verben

Die regelmässigen Verben wurden in folgende Listen aufgeteilt:

- `wordlists/verb-ar.txt` für die Verben wie *gidar – jau gid*, die als regelmässige Verben im engsten Sinn gelten. Diese Liste enthält leider noch Lemmata, die nicht hinein gehören.
- `wordlists/verb-air.txt` für die Verben wie *temair – jau tem*, auch regelmässigen im engsten Sinn.

- `wordlists/verb-er.txt` für die Verben wie *vender – jau vend*, auch regelmässigen im engsten Sinn.
- `wordlists/verb-ir.txt` für die Verben wie *partir – jau part*, auch regelmässigen im engsten Sinn.
- `wordlists/verb-ar-esch.txt` für die Verben wie *gratular – jau gratulesch*, also Verben mit der Endung *-esch* vor den unbetonten Endungen.
- `wordlists/verb-air-esch.txt` für die Verben wie *apparair – jau apparesch*, wobei diese Gruppe sehr klein ist und nicht überall als regelmässig gilt.
- `wordlists/verb-er-esch.txt` für die Verben wie *absolver – jau absolvesch*, auch eine kleine Gruppe und nicht überall als regelmässig gesehen.
- `wordlists/verb-ir-esch.txt` für die Verben wie *finir – jau finesch*, wobei dieser Gruppe viele Lemmata angehören.
- `wordlists/verb-er2.txt` für die Verben wie *currer*, die trotz -er-Endung wie *partir* konjugiert werden. Diese Verben wurden hier implementiert, da sie ohne Aufwand wie die anderen Gruppen verarbeitet werden können.

Nicht als Unregelmässigkeiten zählen die Endung *-el* in der 1. Person Präsens Singular, die Vermeidung von Konsonantenverdoppelungen am Wortende, durch die Schreibweise bedingte Besonderheiten mit ‹c›, ‹g› und ‹gl›, sowie unregelmässige Partizipformen.

Die Endungen (inkl. suffigierte Personalpronomina) für diese Verben sind in `lexc` geschrieben und liegen in folgenden Dateien vor:

- `verb/verb-ar-end.lexc` für *gidar – jau gid*.
- `verb/verb-ar-esch-end.lexc` für *gradular – jau gratulesch*.
- `verb/verb-er-end.lexc` für *temair – jau tem, vender – jau vend*.
- `verb/verb-er-esch-end.lexc` für *apparair – jau apparesch, absolver – jau absolvesch*.
- `verb/verb-ir-end.lexc` für *partir – jau part, currer – jau cur*.
- `verb/verb-ir-esch-end.lexc` für *finir – jau finesch*.

Da der Infinitiv separat implementiert ist, können für verschiedene Verbgruppen die gleichen Endungen verwendet werden. Der richtige Anschluss der Pronomina und die Entscheidung über die Endung *-el* werden durch Ersetzungsregeln in `verb/verb.xfst` sichergestellt.

### 6.12.2 Verben mit Vokalwechsel

Die Verben mit Vokalwechsel weisen in den Formen mit unbetonter Endung einen anderen Stammvokal auf, als in den Formen mit betonter Endung. Auch wenn Regelmässigkeiten existieren, wurde es als einfacher befunden, für jedes Verb beide Stämme zu listen. Diese Verben sind in `verb/verb-vchg.lexc`

implementiert. Für den Anschluss der Pronomina und die richtige Form der Endungen wird auch hier mit Ersetzungsregeln gearbeitet. Zur Regelmässigkeit (abgesehen vom Vokalwechsel) gelten die gleichen Kriterien wie bei den regelmässigen Verben.

### 6.12.3 Unregelmässige Verben

Verben, die nicht in die vorherigen Kategorien passen, gehören zu den unregelmässigen Verben. Diese liegen in der Datei `verb/verb-irr.lexc` in fertiger Form vor. Verben, die sich bloss durch einen Präfix unterscheiden sollten gemeinsam behandelt werden.

### 6.12.4 Unregelmässige Verbpartizipien

Die Partizipformen, die vom allgemeinen Schema abweichen wurden unabhängig von der Konjugationsklasse der Verben in `verb/verb-part-irr.lexc` implementiert. Es muss dabei darauf geachtet werden, nach welchem System (-à, -ì oder konsonantisch) die Partizipien dekliniert werden und ob eine Konsonantenverdoppelung geschieht oder der Stamm auf ‹s› endet und kein ‹s› mehr folgen kann.

Da die unregelmässigen Partizipien die regelmässigen überschreiben müssen parallele Formen hier integriert werden, auch wenn sie regelmässig gebildet würden.

### 6.12.5 Hypothetische Formen

Anhand der Endungen können auch dem System unbekannte Verbformen verarbeitet werden. Dabei werden können sie folgenden Konjugationgruppen angehören:

- Verben wie *gidar – jau gid*.
- Verben wie *temair – jau tem*.
- Verben wie *vender – jau vend*.
- Verben wie *partir – jau part*.
- Verben wie *gratular – jau gratulesch*, allerdings auf gewisse Stammendungen eingeschränkt.
- Verben wie *finir – jau finesch*.

## 7 Schreibregeln

In `spelling/ortho-rule.xfst` sind die Regeln zur Grossschreibung (Erstellung von `fstbinaries/Capitalization.fst`) und die Regeln für die verschiedenen Erscheinungen des Apostrophs und der finalen Verarbeitung der harten und weichen Konsonanten (‹c›, ‹g›, ‹l›; schliesslich in `fstbinaries/OrthoRule.fst`) implementiert.

## 8 Traditionelle Schriftidiome

Für kurze Wörter wie Pronomina, Artikel und einige Präpositionen gibt es pro Idiom in `idioms/` eine `lexc`-Liste, die diese Wörter enthält. Damit können diese Formen, die sich manchmal stark vom Rumantsch Grischun unterscheiden, erkannt werden. Für die sonstigen Fälle sind Ersetzungsregeln für Buchstaben und Buchstabengruppen in `idioms/varieties.xfst` implementiert. Diese können die geläufigsten Lautunterschieden verarbeiten.

Die Transduktoren für die Analyse der Schriftidiome sind nach deren Namen benannt und können auch kombiniert werden. Automatisch erstellt wird die Kombination aus Rumantsch Grischun und den fünf Schriftidiomen.

## 9 Tokenisierung

Das System erwartet Eingabetexte, die grundsätzlich nach Leerstellen tokenisiert wurden. Des weiteren sollten auch Satzzeichen als Tokens stehen, jedoch Zahlen nicht aufgeteilt werden. Mehrworttokens sind nur bei unveränderlichen Wortarten wie Namen erlaubt.

Die Tokenisierung beim Apostroph sollte nach folgender Regel gehen: Ist der Teil vor dem Apostroph verkürzt und nach dem Apostroph ein Vokal, soll getrennt werden und der Apostroph zu ersten Teil gehören (l'onn $\rightarrow$ l' + onn). Ist hingegen der Teil nach dem Apostroph verkürzt und somit ein Konsonant nach dem Apostroph, soll es als ein Token angesehen werden und nicht getrennt werden (gida'l $\rightarrow$ gida'l). Konsonanten hingegen werden im Rätoromanischen nicht durch Apostroph ersetzt.

Ein einfacher Tokeniser, der diese Regeln berücksichtigt ist im Paket enthalten und kann mit `perl` benützt werden:

```
$ perl tokenizer.pl Infile Outfile
```

## 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] Xerox Corporation. Open xerox: Morphological analysis. URL `http://open.xerox.com/Services/fst-nlp-tools/Consume/176` (letzter Zugriff: 2013-07-24). Online-Morphologieanalyse.

[4] Lia Rumantscha. Pledari grond online. URL `http://www.pledarigrond.ch` (letzter Zugriff: 2013-07-07). Onlinewörterbuch für Rumantsch Grischun.