Normalisierung des rätoromanischen Korpus

Dokumentation

In Anlehnung an die Dokumentationen zur Normalisierung des schweizerdeutschen und des französischen Korpus

SNF-sms4science
19.11.2013

Inhaltsverzeichnis

1. Einleitung: Voraussetzungen zur Normalisierung des rätoromanischen Korpus

Das rätoromanische SMS-Korpus setzt sich hinsichtlich der verwendeten Varietäten folgendermassen zusammen:

Rätoromanische Varietät		*Kürzel*	*Anzahl SMS, in denen diese Varietät dominiert ("Hauptsprache")*	*Anteil in %*
Idiome	Sursilvan (Surselva = Bündner Oberland)	roh-sr	426 (davon 1 zu gleichem Anteil Puter[1])	38,0
	Vallader (Unterengadinisch)	roh-vl	336	30,0
	Puter (Oberengadinisch)	roh-pt	181 (davon 1 zu gleichem Anteil Sursilvan)	16,2
	Surmiran (Oberhalbsteinisch)	roh-sm	110	9,8
	Sutsilvan (Sutselva = Imboden, Domleschg, Heinzenberg und Schams)	roh-st	9	0,8
überregionale Schriftsprache Rumantsch Grischun (rg)		roh-gr	59	5,3
		Total	1120	

Die in den meisten SMS benutzten rätoromanischen Idiome (Sursilvan, Vallader, Puter, Surmiran und Sutsilvan) sind regionale Schriftvarietäten, die über in der Volksschule vermittelte Normen verfügen. Innerhalb der angestammten Gebiete dieser Idiome bestehen kleinräumigere regionale sowie lokale gesprochene Varietäten. Als überregionale Schriftsprache für ganz Romanischbünden kommt seit 1982 Rumantsch Grischun zur Anwendung, dessen Normen vor allem aufgrund dreier regionaler Schriftvarietäten (Sursilvan, Surmiran und Vallader) erarbeitet wurden (vgl. Schmid 1982, Darms 1994).

[1] Eine SMS mit derselben Anzahl Wörter in Sursilvan und in Puter ist zweimal verzeichnet.

Bei der vorliegenden Normalisierung von SMS-Daten dient Rumantsch Grischun als Referenzvarietät. Es wird also nicht in die jeweils traditionell zugeordnete Schriftnorm transponiert, sondern – sowohl aus kleinräumigeren, spontan verschriftlichten Varietäten als auch aus den traditionellen regionalen Schriftvarietäten – in die überregionale Schriftsprache Rumantsch Grischun. Die Normalisierung richtet sich vor allem nach dem Online-Wörterbuch *Pledari Grond* (www.pledarigrond.ch, Stand Juni 2012[2]). Für spezifische Aspekte werden auch Wörterbücher der Idiome herangezogen:

Rätoromanische Varietät		*Kürzel*	*Referenzwörterbuch*
Idiome	Sursilvan (Surselva = Bündner Oberland)	roh-sr	Decurtins 2001
	Vallader (Unterengadinisch)	roh-vl	Tscharner 2009
	Puter (Oberengadinisch)	roh-pt	Tscharner 2007
	Surmiran (Oberhalbsteinisch)	roh-sm	Signorell 1999
	Sutsilvan (Sutselva = Imboden, Domleschg, Heinzenberg und Schams)	roh-st	Eichenhofer 2009
überregionale Schriftsprache Rumantsch Grischun (gr)		roh-gr	*Pledari Grond*

Die Normalisierung beschränkt sich vorerst auf eine Wort-für-Wort-Umsetzung, klammert also Syntax und Idiomatik aus. Ziel dieser Umsetzung, die hier als Glossierung bezeichnet wird, ist ein Layer, der durchsuchbar ist, weil er eine standardisierte Graphie verwendet.

Die Glossierung in Rumantsch Grischun ist im Hinblick auf das Lexikon mit einer Reihe von Entscheiden verbunden. Oft haben die verschiedenen Ausgangsvarietäten – je nachdem auch die Zielvarietät – nicht denselben Worttyp bzw. wenn, dann mit unterschiedlicher Bedeutung. Ein weiteres Problem stellt der Umgang mit anderssprachigem Material dar. Entlehnungen können einerseits im

[2] Da dieses Wörterbuch ständig aktualisiert wird, lassen sich Entscheide bei der Glossierung, die aufgrund des Standes vom Juni 2012 getroffen wurden, unter Umständen später nicht mehr nachvollziehen.

Referenzwörterbuch einer bzw. in den Referenzwörterbüchern mehrerer regionaler Schriftsprachen verzeichnet sein, im *Pledari Grond* aber fehlen. Andererseits gibt es Entlehnungen, die ins *Pledari Grond*, jedoch nicht in die regionalsprachlichen Normwerke Eingang gefunden haben.

Die Prinzipien, die der Normalisierung der lexikalisch erheblich divergierenden rätoromanischen Varietäten zu Grunde liegen, werden in dieser Dokumentation erläutert. Ausserdem werden allgemeine Normalisierungsregeln, die sich nicht nur auf die rätoromanischen Daten beziehen, mit Beispielen veranschaulicht.

Alle Einträge des *Pledari Grond* und der ausgewählten regionalsprachlichen Wörterbücher gelten – unabhängig von varietätenspezifischen Zuweisungen (wie z.B. im *Pledari Grond*: *fam.* = familiär, *giuv.* = *giuvenils* 'jugendsprachlich' oder – bei diatopischer Markierung – *S.* = *Surselva*, *C.* = *Grischun central* 'Mittelbünden', *E.* = *Engiadina*) und ohne Rücksicht auf abweichende Semantik als Referenzangaben für die unten in Kapitel 3 erläuterten Glossierungsregeln.

2. Tokens

Durch ein automatisches Betriebssystem werden im *sms4science*-Korpus einzelne Tokens voneinander unterschieden. Ein Token entspricht nicht in jedem Fall einem Wort, es handelt sich um durch Leerzeichen, Apostrophe oder Bindestriche getrennte Elemente. Ausserdem wird auch die Mehrzahl der Satzzeichen und Smileys durch die Software erkannt und als Tokens dargestellt. Es ist möglich, dass die Anzahl der Tokens in einer SMS nicht der Anzahl der Tokens in der normalisierten Fassung entspricht. Ausserdem ist darauf hinzuweisen, dass das Betriebssystem die Tokens nicht immer korrekt identifiziert. Solche Fehler müssen von Hand korrigiert werden.

2.1. Ein Token in der SMS – mehrere Tokens in der Normalisierung

In einer SMS kann ein automatisch identifiziertes Token in Wirklichkeit mehreren Tokens entsprechen. Dies ist zum Beispiel bei Akronymen wie *Bn* (= *Buna notg*, "Gute

Nacht") oder bei agglutinierten enklitischen Subjektpronomen wie in Sursilvan *possu* (= *poss(el) jeu* → RG[3] *poss jau*, "mag ich") der Fall.

Einem Token (*possu*) entsprechen somit zwei normalisierte Tokens (*poss* | *jau*). Um dies auf der glossierten Ebene festhalten zu können, müssen im selben Kästchen unter dem Originaltoken zwei durch einen Leerschlag getrennte Wörter eingefügt werden.

2.2. Mehrere Tokens in der SMS – ein Token in der Normalisierung

Der Fall, in dem mehrere Tokens (z.B. *o* | *.* | *k* oder ca | .) in der Normalisierung zusammen nur ein Token (*okay*, *circa*) ergeben, kommt im rätoromanischen Korpus seltener vor. In solchen Fällen wird nur das erste Kästchen ausgefüllt, die anderen bleiben leer.

2.3. Falsche Tokenisierung

Bei der automatischen Tokenisierung der SMS werden Sequenzen manchmal fehlerhaft in Tokens unterteilt. Ein Beispiel ist die Unterteilung von *n8* in die Tokens *n* und *8*. Hierbei handelt es sich lediglich um ein Token (*notg*), da die beiden Zeichen nur eine lexikalische Einheit darstellen (*n8 = notg* 'Abend'). Auch Apostrophe sind häufig nicht korrekt tokenisiert, da sie automatisch als Satzzeichen (PUN) gekennzeichnet werden. Der Ausdruck *in'autra* ("eine andere") wird somit in drei Tokens geteilt (*in* | ' | *autra*). Um dies zu korrigieren, wählt man im Teil "SMS Details" unter Status "needs retokenizing" an und vermerkt, dass die Tokens *in* und *'* zu einem einzigen Token zusammengefasst werden müssen. Die falschen Tokenisierungen werden in einem späteren Schritt von Hand korrigiert.

[3] RG = Rumantsch Grischun.

2.4. Zusätzliche Informationen

Unter den beiden Hauptlinien auf der Normalisierungsseite (Nachricht und Glossierung) ist eine weitere Linie mit Zusatzinformationen zu sehen. Dabei handelt es sich um die als *PoS* (*Part of Speech*) bezeichnete Linie, worin hauptsächlich automatisch generierte Aspekte wie die Zeichensetzung (PUN), lautmalerische Elemente wie *haha* (ONO), Zahlen (NUM) und die Emoticons (EMO) normalisiert sind. Wenn in diesen Feldern Fehler vorkommen, muss dies im Kästchen *notes* angemerkt werden.

Spezialfall:

In den Idiomen Vallader und Puter schliesst die Form der 1. Person Singular Präsens[4] des Verbs *avair* 'haben' eine agglutinierte, in der heutigen Sprache nicht mehr motivierte Partikel ein, die nach der orthographischen Norm weiterhin als separates Element dargestellt wird: Vallader *eu n'ha* bzw. Puter *eau d'he* ("ich habe") sowie bei Inversion *n'haja* bzw. *d'heja* ("habe ich")[5].

Neben von gewissen Personen bevorzugten oder sporadisch verwendeten Graphien (Vallader *nha(ja)* statt *n'ha(ja)*, SMS 24006, 25255; Puter *dhe* und *dä* statt *d'he*, SMS 7760, 24052) wird im Korpus v.a. die der Norm entsprechende Graphie mit Apostroph verwendet: *N'ha quist nouv nr. da handy* ("Habe diese neue handynr.", SMS 18519), *d'he be güst gieu temp* ("habe nur gerade zeit gehabt", SMS 8286). In diesen Fällen erfolgte eine falsche Tokenisierung der Original-SMS (*n* | *'* | *ha*, *n* | *'* | *haja* bzw. *d* | *'* | *he*, *d* | *'* | *heja*). Bei der erforderlichen Retokenisierung sind die beiden durch einen Apostroph getrennten Komponenten einem einzigen Token zuzuweisen.

Homograph zu Vallader *n'haja* ("habe ich") ist Rumantsch Grischun *n'haja* in der Verbindung *n'haja betg* ("habe nicht", 1./3. Person Singular Konjunktiv Präsens), wo *n'* präverbale Negationspartikel ist und somit ein eigenes Token darstellt.

[4] Dies gilt sowohl für den Indikativ als auch für den Konjunktiv. Da Letzterer im Korpus nicht belegt ist, wird er hier nicht berücksichtigt.

[5] Sowohl *n'* als auch *d'* sind als Nachfolger von lat. INDE erklärbar (daraus auch fr. *en* und it. *ne* 'davon').

3. Idiome – Rumantsch Grischun: Schnittstellen und Unterschiede

Die Normalisierung der rätoromanischen Daten nimmt wie erwähnt in erster Linie auf das Online-Wörterbuch *Pledari Grond* (Rumantsch Grischun) Bezug. Wegen der Verschiedenheit der fünf Idiome kann es jedoch vorkommen, dass sich ein Worttyp nicht im *Pledari Grond* , sondern nur in regionalsprachlichen Normwerken findet bzw. dass ein Worttyp im *Pledari Grond* mit anderer Semantik als in den regionalsprachlichen Normwerken verzeichnet ist.

Die Frage vor der Normalisierung jedes Tokens lautet:

Existiert das verwendete Wort in Rumantsch Grischun?

Damit ist das Vorkommen des Wortes mit Berücksichtigung aller orthographischen Varianten gemeint (z.B. RG/Vallader *chasa*, Puter *chesa*, Surmiran *tgesa*, Sutsilvan *tgea*, Sursilvan *casa*).

Wenn die Antwort auf diese Frage ja lautet, wird gemäss 3.1 vorgegangen. Ist das Wort nicht im *Pledari Grond* verzeichnet, wird gemäss 3.2 normalisiert.

3.1. Wörter aus Idiomen mit Äquivalent in Rumantsch Grischun

Wenn ein lexikalischer Typ im *Pledari Grond* vertreten ist, wird er beibehalten, dies unabhängig von semantischen Unterschieden zwischen Ausgangs- und Zielvarietät.

In den folgenden Beispielen stimmen Worttyp und Bedeutung in der Ausgangs- und in der Zielvarietät überein:

(1) Puter *chesa* ‘Haus’ → RG *chasa*
(2) Sursilvan *tier* ‘bei, zu’ → RG *tar*
(3) Surmiran *nous* ‘wir, uns’ → RG *nus*

Eine andere Möglichkeit beim Auftreten von Äquivalenten ist die, dass zwar der Worttyp übereinstimmt, die Bedeutung aber nicht oder nicht vollumfänglich. Der Einfachheit halber werden diese Wörter trotzdem anhand der Äquivalente in Rumantsch Grischun

normalisiert, damit an der Originalnachricht nicht zu grosse Veränderungen vorzunehmen sind:

(4) Sursilvan *migliar* ‘essen, fressen’ → RG *magliar* ‘fressen’
(5) Vallader *massa* ‘zu viel’ → RG *massa* ‘Masse’
(6) Sursilvan/Münstertalisch *schon* ‘bereits’ → RG *schon* ‘doch, wirklich’

Treten bei der Arbeit mit der normalisierten Version Unklarheiten auf, ist auf die nicht normalisierte Version und auf die Wörterbücher der jeweiligen Idiome zurückzugreifen.

3.2. Wörter aus Idiomen ohne Äquivalent in Rumantsch Grischun

Aus Idiomen stammende Wörter, die in Rumantsch Grischun keine Entsprechung innerhalb des Worttyps haben, müssen ins Rumantsch Grischun übersetzt werden:

(7) Vallader/Puter *dafatta* ‘sogar’ → RG *perfin*
(8) Vallader/Puter *darcheu/darcho* ‘wieder’ → RG *puspè*
(9) Sursilvan *ni* ‘oder’ → RG *u*

Übersetzt werden auch Entlehnungen, die nur im jeweiligen Idiom-Wörterbuch, nicht jedoch im *Pledari Grond* verzeichnet sind. Das hat zur Folge, dass anderssprachiges Wortgut in der normalisierten Version nicht mehr erkennbar ist. Diese Lösung ist aber doch die praktikabelste im Umgang mit den unterschiedlichen Integrationsgraden anderssprachiger Wörter in den verschiedenen Idiomen. Integrierte (d.h. im jeweiligen Idiom-Wörterbuch registrierte) Entlehnungen werden als rätoromanische Wörter behandelt und auf Rumantsch Grischun übersetzt, nicht integrierte als Nonce-Borrowings, bei denen die Herkunft vermerkt wird:

(10) Sursilvan *also* → RG *pia*
im Gegensatz zu Vallader/Puter *also* → *also* [DE]
(11) Sursilvan *aber* → RG *però*
(12) Puter *apunto*, Vallader *apunta* ‘eben’ → RG *eba*

4. Gross- und Kleinschreibung

Die Ortographie der normalisierten Tokens richtet sich nach dem jeweiligen Eintrag im *Pledari Grond*. Daraus ergibt sich, dass in der normalisierten Fassung am Satzanfang häufig klein geschriebene Wörter stehen. Andere Wörter wie *Dieus* 'Gott' (RG *Dieu*) werden glossiert gross geschrieben, obwohl sie in der Originalnachricht klein geschrieben sind.

Gemäss der Orthographienorm des Rumantsch Grischun werden einfache Eigennamen, geographische Bezeichnungen, Feiertagsbezeichnungen, Gott- und Heiligenbezeichnungen sowie Länderbezeichnungen gross geschrieben. Adjektive, die Nationalitäten- und Sprachzugehörigkeiten betreffen, werden klein geschrieben (*tudestg* 'deutsch', *rumantsch* 'rätoromanisch'); sind sie jedoch substantiviert und dienen sie der Bezeichnung von Personen, wird gross geschrieben (*Rumantsch*[*a*] 'Rätoromane, -in'). Weiter sind Wörter in Titelbezeichnungen sowie in gewissen Abkürzungen gross geschrieben.

5. Apostroph

Wortformen, bei denen in den rätoromanischen Originalnachrichten Laute elidiert sind (Markierung mit Apostroph), werden in der normalisierten Fassung ausgeschrieben.

Einen Spezialfall stellt der maskuline Artikel im Singular dar: Die nach der Norm des Rumantsch Grischun vor vokalischem Anlaut zu verwendende Variante mit Apostroph (*l'*) wird fallen gelassen, zugunsten der nach der Norm nur vor konsonantischem Anlaut zu verwendenden Variante *il* (cf. Bsp. 16).

Die Regel zum Auschreiben elidierter Wortformen erstreckt sich nicht auf den oben unter 2.4 erwähnten Spezialfall *n'ha/d'he* '(ich) habe', wo keine in der heutigen Sprache nachvollziehbare Elision vorliegt.

Die in den folgenden Beispielen Wort um Wort umgesetzten Sequenzen entsprechen in den meisten Fällen nicht der Norm des Rumantsch Grischun, welche die Elision vorschreibt (eine Ausnahme bildet Bsp. 15, wo die Elision fakultativ ist).

(13) Sursilvan *in'autra* ("eine andere") → RG *ina autra*

(14) Vallader *ün'eivna* ("eine Woche") → RG *ina emna*

(15) Vallader/Puter/Sursilvan *ch'el*(*s*) ("dass er/sie$_{Pl.}$", Pron. Rel. + er/sie $_{Pl.}$)
→ RG *che el*(*s*)

(16) Vallader/Puter *l'aviun* ("das Flugzeug") → RG *il aviun*

(17) Vallader/Puter/Sursilvan *l'energia* → RG *la energia*

(18) Sursilvan *dall'ina* ("um ein Uhr") → RG *da la ina*

Auch bei nicht integrierten anderssprachigen Ausdrücken (Nonce-Borrowing) werden elidierte Wortformen ausgeschrieben, so im italienischen *Sogni d'oro*, das zu *sogni di oro* normalisiert wird.

6. Ellipsen

Bei Ellipsen in der Originalnachricht werden in der normalisierten Fassung keine Ausdrücke ergänzt. Im folgenden Beispiel wird also kein Pronomen hinzugefügt, auch wenn das fehlende Element eindeutig rekonstruierbar ist:

(19) *Sun en iert* ("Bin im Garten") wird nicht zu *jeu sun en iert* ("ich bin im Garten")

Motivation

Jede Art von Ellipse kann für Sprachforschungszwecke interessant sein. Dieser Aspekt darf deshalb bei der Normalisierung nicht verloren gehen.

7. Graphische Varianten

Grundsätzlich werden alle graphischen Varianten normalisiert, das heisst der im *Pledari Grond* aufgeführten Form angepasst:

(20) Puter *vainst eeeeir?* (“kommst auch?”) → RG *vegns er?*
(21) Puter *Graaazcha* (‘Danke’) → RG *grazia*
(22) Sursilvan *saaaau*, *tschauiii* (‘ciao’) → RG *tgau*

Dies gilt auch für Fälle, die sich in ihrer graphischen Variation einer anderssprachigen Graphie angleichen (23). Ausserdem wird diese Regel für alle als anderssprachig bezeichneten Ausdrücke (24, 25) angewendet:

(23) *kaputta* → RG *caputta*
(24) *Geeelateriaaaaa* → ital. *gelateria*
(25) *häppy börsdei* → engl. *happy birthday*

Eine Ausnahme von dieser Regel gilt für Onomatopoetika wie *muaah*. Zur Einteilung dieser Ausdrücke wird auf Kapitel 10 verwiesen.

8. Zusammen- und Getrenntschreibung bei Personal- und Reflexivpronomen, Präpositionen und Artikeln

Das Subjektpronomen wird in enklitischer Position – gemäss der vorherrschenden Gebrauchsnorm des Rumantsch Grischun und im Gegensatz zur vorherrschenden Gebrauchsnorm des Vallader und des Puter – in seiner Vollform verwendet und getrennt vom Verb geschrieben (26, 27), mit Ausnahme des areferentiellen Pronomens *i* ‘es’, das – gemäss der Norm des Rumantsch Grischun – an die Verbform agglutiniert wird (28).

Im Gegensatz zu dem im Sursilvan nicht flektierten und an die Verbform agglutinierten Reflexivum wird das Reflexivum in der normalisierten Fassung, gemäss der Norm des Rumantsch Grischun, flektiert und getrennt von der Verbform geschrieben (29).

Artikel werden gemäss der Norm des Rumantsch Grischun nur mit den Präpositionen *a* und *da* verschmolzen, und dies nur bei den maskulinen Formen: *al(s)*, *dal(s)*. Beim femininen Artikel in Verbindung mit *a* und *da* sowie bei den übrigen Präpositionen wird getrennt geschrieben (30).

Einen Spezialfall stellt die Fusion der Präpositionen *vi* und *da* zu *vid* im Surselvischen dar, der im Rumantsch Grischun die Verbindung *vi da* entspricht (31).

(26) Vallader *possa* (“kann ich”) → RG *poss jau*
(27) Vallader *sto’la* (“muss sie”) → RG *sto ella*
(28) Sursilvan *datti* (“gibt es”) → RG *datti*
(29) Sursilvan *selegrel* (“freue mich”) → RG *ma legrel*
(30) Sursilvan *ella maschina* (“in der Maschine”) → RG *en la maschina*
(31) Sursilvan *vid* (‘an’) → RG *vi da*

Die elidierten (27) und agglutinierten (28) Subjektpronomen werden nicht als Abkürzungen markiert.

9. Komposita

Komposita, die aus mehr als einem Token (Wort) bestehen, müssen als lexikalische Einheit gekennzeichnet werden. Dafür werden alle Tokens in das erste Feld geschrieben, während die weiteren Felder leer bleiben.

(31) RG *grippa* | *da* | *portgs* ‘Schweinegrippe’ → RG *grippa da portgs* | | |

Es werden nur Komposita als solche berücksichtigt, die im *Pledari Grond* verzeichnet sind.

10. Abkürzungen

10.1. Phraseologische Abkürzungen

In SMS werden oft Kurzschreibungen verwendet. Dabei handelt es sich sowohl um Kurzschreibungen, die man in grossen Teilen der Sprechergruppe kennt, wie *cs* für *cars salids* ("liebe Grüsse"), als auch um Kurzschreibungen, die nur in kleinen Gruppen oder sogar nur zwischen dem Sender und dem Empfänger bekannt sind. Im Korpus möchten wir gerne alle Kurzschreibungen als solche kennzeichnen, damit auch danach gesucht werden kann. Deshalb wird jedes Token, das eine Kurzschreibung darstellt, als "Abbreviation" (A) markiert. Akronyme (wie *cseb* für *chars salids e bitsch*, "liebe Grüsse und Kuss") und konventionelle Abkürzungen (*evtl.* für *eventualmein* 'eventuell') werden dabei gleich kategorisiert, um den Aufwand in Grenzen zu halten. Sollte sich in Zukunft jemand wissenschaftlich für Kurzformen interessieren, so kann diese Person nach allen Formen suchen, die als "Abbreviation" markiert sind, und dann eigene Klassifizierungen vornehmen.

Nicht standardisierte (d.h. nicht im *Pledari Grond* verzeichnete) Kurzschreibungen, die allgemein bekannt oder offensichtlich erkennbar sind, werden auf der glossierten Ebene ausgeschrieben. Solche Kurzschreibungen werden ebenfalls als "Abbreviation" markiert.

10.2. Abgekürzte Eigennamen, Akronyme, lexikalisierte Kurzschreibungen

Neben den oben erwähnten Kurzschreibungen, die auf Phrasen basieren, gibt es Kurzformen. Dazu gehören *ETH* für *Eidgenössische Technische Hochschule*, *Laser* für *Light Amplification by Stimulated Emission of Radiation*, aber auch *VS* für *Vallais* oder *du* für *dumengia*. Bei der Entscheidung, welche dieser Kurzschreibungen auf der glossierten Ebene ausgeschrieben werden sollen und welche nicht, lassen wir uns von zwei Fragen leiten:

1) Was ist zumutbar für die Mitarbeiter, die glossieren?

2) Welche Formen können unter Umständen auf Interesse bei Forschenden stossen?

Daraus ergeben sich folgende Regeln:

1. Formen, die in abgekürzter Form im *Pledari Grond* zu finden sind, werden auf der glossierten Ebene nicht analysiert (*Laser* bleibt *Laser*).
2. Eigennamen werden nicht analysiert (*ETH* bleibt *ETH*).
3. Anders sieht es beim folgendem Beispiel aus:
   Va bää;) ***m!*** *fluriii eau vaiva pers il natel […]* ("Ist in ordnung ;) m! fluriii ich hatte das natel vergessen […]")
   Hier können wir nicht wissen, für was der Buchstabe *m* steht. Dementsprechend belassen wir auf der glossierten Ebene *m* und markieren dieses Token zusätzlich als "unclear".
4. Abgekürzte Personennamen, wie z.B. in *da cor,* ***a*** ("von herzen, **a**") werden nicht als "Abbreviation" markiert, weil sie für die Forschung nicht interessant sind.
5. Kurzformen, die nur aus Zeichen bestehen, werden auch als Abkürzung markiert und ausgeschrieben (z.B. *+* → *e*). Falls sie als Zeichensetzung (PUN) markiert waren, muss die Nachricht mit dem Vermerk "needs retokenizing" versehen werden.

Beispiele:

Wenn in der Originalnachricht *evtl.* (für *eventualmein*) steht, wird diese Form als Abkürzung gekennzeichnet und ihrem Äquivalent im Wörterbuch (*ev.*) angepasst.

(32) *evtl.* → *ev.*

Anders wird bei der Abkürzung *ca.* vorgegangen. Da diese – obwohl es sich um eine konventionelle Abkürzung handelt – nicht im *Pledari Grond* steht, wird sie in allen Fällen ausgeschrieben. Bei Abkürzungen mit Punkt muss darauf geachtet werden, dass der Punkt nicht als Satzzeichen (PUN), sondern als zum vorausgehenden Token (dem

abgekürzten Element) zugehörig markiert ist. Wenn dies nicht der Fall ist, muss “needs retokenizing” angeklickt werden, damit eine Neuzuordnung erfolgt.

(33) *ca.* → *circa*

Einen Spezialfall stellt die nur anderssprachig motivierbare (fr. *heure(s)*), jedoch rätoromanisch “gelesene” Abkürzung *h* dar, die kein Zeichen des gemeinten rätoromanischen Wortes enthält. Sie wird in der normalisierten Version mit dem Wort *uras* ‘Stunden’ ausgeschrieben und mit dem Attribut “Abbreviation” versehen.

(34) *17 h* → *17 uras*

10.3. Anderssprachige Abkürzungen

Auch anderssprachige Abkürzungen erhalten (neben der Angabe der Sprache) die Bezeichnung “Abbreviation”. Hier werden – unabhängig von den anderssprachigen Wörterbüchern – alle Abkürzungen ausgeschrieben (mit Ausnahme der Namen, vgl. Beispiel *ETH*).

(35) *cu* oder *cya* → *see you*
(36) *sry* → *sorry*

Auch Abkürzungen, die nur einzelne anderssprachige Elemente enthalten, werden als anderssprachig markiert. Das Akronym *jammmbt* wird zu *jau hai mega mega mega gugent tai* (“ich habe dich mega mega mega gern”) und als schweizerdeutsch gekennzeichnet, damit die Information, dass die Insertion *mega mega mega* schweizerdeutsch ist, nicht verloren geht.

10.4. Rebusschreibung

Auch Rebusschreibungen – im rätoromanischen Korpus die Verwendung von Zahlzeichen für Sequenzen, die mit dem entsprechenden Zahlwort homophon sind[6] – werden als Abkürzungen behandelt.

(37) *buna **n8*** → *buna notg*

Sie werden ausgeschrieben und als Abkürzung markiert. Wenn die Zahl durch die automatische Erkennung als NUM bezeichnet wurde, muss als Bemerkung "needs retokenizing" angegeben werden, damit die Elemente *n* und *8* als ein einziges Token angesehen werden.

11. Namen

Das Korpus ist bereits anonymisiert. Eigennamen können also in aller Regel nicht mehr den Schreibern oder deren Bekannten zugeordnet werden. Nachnamen wurden ersetzt durch <Lastname> und Vornamen wurden "rotiert", d.h. ein Vorname wurde durch einen andern ersetzt. Diese Anonymisierung ist jedoch nicht vollständig. Dies einerseits, weil gewisse Namen nicht anonymisiert werden können, da sie homograph zu Einheiten des im Korpus vertretenen Lexikons sind (die weiblichen Vornamen *Pia* und *Adina* z.B. werden gleich geschrieben wie die surselvischen Adverbien *pia* 'also' und *adina* 'immer', der männliche Vorname *Aschi* [schweizerdeutsche Variante zu *Ernst*] wird gleich geschrieben wie das surselvische Adverb *aschi* 'so'). Andererseits wurden aber auch einige Namen nicht bearbeitet. Da nun alle SMS nochmals überprüft werden, ist eine gute Gelegenheit gegeben, die Anonymisierung zu kontrollieren.

[6] Nicht belegt ist die Verwendung von Buchstaben für Sequenzen, die mit der entsprechenden Buchstabenbezeichnung homophon sind, also etwa *b* (gesprochen [be]) für *be* 'nur'.

11.1. Woran erkennt man, ob ein Name anonymisiert ist?

Namen, die anonymisiert sind, sind im System als solche markiert. Entsprechend wurden sie beim Erstellen des Arbeitskorpus gleich in die zweite Zeile ("Gloss") kopiert. Findet sich also in dieser Zeile bereits der gleiche Name wie in der Zeile "Message", besteht kein Handlungsbedarf. Steht jedoch in der zweiten Zeile kein Name, so sollte man die SMS mit dem Attribut "Anonymisation" markieren. Der Name muss danach nicht auf die glossierte Ebene übernommen werden, da er rotiert wird. Dies gilt auch für Namen, die offensichtlich nicht anonymisiert werden können wie *Pia* oder *Adina*, denn diese Markierung wird in der computerlinguistischen Verarbeitung gebraucht.

11.2. Abgekürzte Namen

Abgekürzte Namen werden nicht als "Abbreviation" markiert (vgl. Kapitel 10.2).

11.3. Produktenamen

Produktenamen wie z.B. Facebook werden nicht anonymisiert und auch nicht markiert. Sie werden auf der glossierten Ebene gegebenenfalls orthographisch angepasst (*faceebbookk* → Facebook) aber ansonsten ganz übernommen.

11.4. Toponyme

Bei Toponymen wird jedes Wort als einzelnes Token angesehen. Der Ausdruck *Lac Lémand* besteht also aus zwei Tokens. Weil wir Eigennamen bei der Annotation der Sprachen ausschliessen, wird hier *Lac* nicht als französische Entlehnung angesehen. Die einzelnen Tokens werden ganz auf die glossierte Ebene übernommen und nur orthographisch angepasst (*Lemand* → *Léman*).

(38) *Lac Lemand* → *Lac Léman*

12. Onomatopoetika

Ähnlich wie in Comics werden in SMS oft lautmalerische Formen wie *phuu*, *whoaa* verwendet. Diese Formen müssen markiert werden, damit einerseits danach gesucht werden kann. Andererseits stellen sie, wenn sie markiert sind, kein Problem für den PoS-Tagger und andere Software-Tools dar.

12.1. Abgrenzung

In der Linguistik greift der Begriff *Onomatopoetikum* sehr weit. Er deckt auch Wörter wie *cucu* 'Kuckuck' oder *miau*(*l*)*ar* 'miauen' ab. Derartige Lexeme wollen wir aber nicht markieren. Die Abgrenzung des hier zu verwendenden Verständnisses von Onomatopoetikum ist also schwierig. Der Übergang zu Lexemen wie *grir* 'schreien' einerseits und Interjektionen wie *ah* andererseits ist fliessend. Der hier verwendete Ansatz soll deshalb vor allem praktikabel sein und sich an der Frage orientieren, was mit der Markierung dieser Tokens letztlich bezweckt wird.

Die Markierung dieser Tokens verfolgt zwei Ziele. Einerseits sollen die Tokens als etwas markiert sein, das beim Part-of-Speech-Tagging ignoriert werden kann. Andererseits sollen sie als Gruppe für zukünftige Forschungsprojekte auffindbar sein. Beide Ziele werden am besten mit folgenden Einschränkungen des Begriffs *Onomatopoetikum* erreicht:

- Tokens, die in der gegebenen Schreibweise nicht im *Pledari Grond* enthalten sind.
- Tokens, die einen lautmalerischen Charakter haben.

Mit dieser Definition weichen wir stark vom üblichen Gebrauch des Begriffs ab. Dies nehmen wir in Kauf, denn wir erreichen so die gesetzten Ziele am besten:

- Alle derart markieren Tokens können von computerlinguistischen Analysen ausgeschlossen werden.
- Tokens, die der Computerlinguistik keine Probleme bereiten, weil sie in Wörterbüchern enthalten sind (z.B. *ah!*), bereiten uns Mitarbeitern während der Glossierung keinen zusätzlichen Aufwand.

Im resultierenden Wörterbuch, das alle Tokens/Types/Lemmata enthält, können die derart markierten Tokens leicht mit andern Onomatopoetika wie *ah* zusammengefasst werden, falls dies gewünscht ist, denn die Formen, die den Kodizes entsprechend geschrieben sind, können im Wörterbuch leicht gefunden und gruppiert werden.
Wir weichen mit diesem Vorgehen aber von einem anderen Prinzip ab, dem wir beim Arbeiten mit dem Korpus üblicherweise folgen: Wir beziehen ausnahmsweise die Graphie in unsere Analyse mit ein. Auch diese Inkonsequenz nehmen wir in Kauf, weil wir klare Ziele vor Augen haben, die wir mit möglichst geringem Aufwand erreichen wollen.

12.2. Konkretes Vorgehen

Konkret werden Tokens mit lautmalerischem Charakter als erstes im *Pledari Grond* gesucht. Sind sie dort zu finden (z.B. *ah*), so werden sie so (u.U. angepasst an die Gross-/Kleinschreibung) auf die glossierte Ebene übernommen. Findet man sie nicht im *Pledari Grond* (z.B. *phuu* und *whoaa*), so werden sie auf die glossierte Ebene übernommen und als Onomatopoetika markiert.

13. Orthographie

Die Orthographie richtet sich nach dem *Pledari Grond*. Bei der Normalisierung von anderssprachigen Elementen werden die Nachschlagewerke in den jeweiligen Sprachen konsultiert.

14. Sprachen

Die SMS wurden bereits nach Sprachen annotiert, d.h. es wurde für jede SMS bestimmt, welches die Hauptsprache ist und welche Entlehnungen und Ad-hoc-Entlehnungen sich darin finden lassen. Diese Bestimmung fand pro SMS statt, die einzelnen Wörter haben also keine Sprachmarkierung. Diesen Schritt möchten wir nun für Ad-hoc-Entlehnungen

nachholen. Das Token *cu* wird also als englisch annotiert. Das im *Pledari Grond* verzeichnete *café* hingegen wird nicht als Entlehnung markiert.

14.1. Berücksichtigte Sprachen

Bei der Spracherkennung auf der Ebene SMS wurden diverse Sprachen annotiert. Davon ausgehend, dass Untersuchungen am Korpus in den Landessprachen oder in Englisch ausgeführt werden, werden nur diese Sprachen berücksichtigt. Ad-hoc-Entlehnungen aus andern Sprachen werden als “other” markiert. Anders als bei der Bestimmung der Hauptsprachen der Nachrichten, wo nach fünf Idiomen und Rumantsch Grischun differenziert wurde, werden bei der Zuweisung der Ad-hoc-Entlehnungen keine rätoromanischen Varietäten unterschieden, sondern es erfolgt nur eine Einstufung als Rätoromanisch. Dem Gebrauch verschiedener Idiome innerhalb einer Nachricht kann somit nicht Rechnung getragen werden.

14.2. Praktisches Vorgehen

Wird ein Token z.B. als englisch markiert, so kopiert das System dieses direkt in die zweite Zeile. Praktischerweise bestimmt man bei Ad-hoc-Entlehnungen also zuerst die Sprache.

14.3. Anpassen der Orthographie

SMS weisen manchmal eine lautnahe Schreibweise auf, d.h. sie sind oft pseudophonetisch geschrieben. So kann z.B. *cool* als *cuul* auftreten. Derartige Schreibweisen werden auf der glossierten Ebene wenn immer möglich korrigiert. Steht auf der SMS-Ebene also *cuul*, so wird dies auf der glossierten Ebene zu *cool*.

14.4. Anpassen der Flexion

Anderssprachige Tokens werden oft morphologisch ans Rätoromanische angepasst. So könnte sich beispielsweise folgende Form in einer SMS befinden: *ina cuula caussa*. Wie bereits gesagt, würde hier die englische Schreibweise verwendet. Zusätzlich wird diese aber – wenn sie bereits im Originaltext morphologisch angepasst war – auch in der normalisierten Fassung entsprechend angepasst. Wenn nicht klar ist, ob ein anderssprachiges Token an die rätoromanische Morphologie angepasst ist (z.B. *relaxantas*) oder nicht (z.B. *cool*), so wird folgendermassen evaluiert:

- Ist das Token im *Pledari Grond* zu finden? Wenn dem so ist, so wird davon ausgegangen, dass auch die Flexion dem romanischen Muster folgt, ausser in Fällen, in denen das Wörterbuch etwas anderes angibt. Die Angaben des *Pledari Grond* haben also oberste Priorität. Das Token ist in dem Fall auch keine Ad-hoc-Entlehnung, die Sprache wird also nicht markiert.
- Wenn das Token nicht im *Pledari Grond* steht, so wird die Flexionsform übernommen, die vom Schreiber verwendet wird, sofern sie entweder der Ausgangssprache oder dem Rätoromanischen entspricht.
- Entspricht die Flexionsform weder der Ausgangssprache noch dem Rätoromanischen, so wird sie ans Rätoromanische angepasst.

14.5. Anpassen der Gross-/Kleinschreibung

Ad-hoc-Entlehnungen aus dem Deutschen, die substantivische Funktion haben, werden gross geschrieben.

14.6. Toponyme und Eigennamen

Toponyme und Eigennamen in fremden Sprachen (z.B. *Ticino*, *New York*, *IBM*) werden Eins zu Eins übernommen und nicht als Ad-hoc-Entlehnungen markiert. Dies, weil es unmöglich ist, eine Abgrenzung vorzunehmen. *Paris* beispielsweise ist sowohl die rätoromanische als auch die französische Form des Namens. Andererseits müsste man

sonst *IBM* tatsächlich als englische Entlehnung markieren, denn das Akronym steht für *International Business Machines*. Dieses Vorgehen ist im Hinblick auf mögliche weitere Untersuchungen weder zumutbar noch sinnvoll.

15. Unclear

In der Glossierungsmaske gibt es die Auswahl “unclear”. Diese wird für Formen gebraucht, bei denen nicht gesagt werden kann, was der SMS-Schreiber meinte. Es kann sich dabei um Tippfehler handeln, die so gravierend sind, dass der Sinn nicht erkannt werden kann, oder um Wörter in unbekannten Sprachen oder um homophone Formen, bei denen auch im Syntagma nicht klar wird, was gemeint ist. Alle diese Formen werden so belassen, wie sie in der SMS stehen (auch die Gross-/Kleinschreibung wird nicht verändert!) und als “unclear” markiert.

16. Zahlen

Zahlen werden so belassen, wie sie im Text auftauchen. Wenn Zahlen als Rebusschreibung verwendet werden, so werden sie übersetzt: *n8* wird folglich zu *notg* ‘Nacht’. Treten sie in Kombination mit Abkürzungen auf (z.B. im Falle von *2x*), so werden die Zahlen so belassen. Die Abkürzung wird aber umgesetzt: *2 giadas* (“2 Mal”).

17. Zusammenschreibung

Grundsätzlich gilt die Schreibweise des *Pledari Grond*. Wenn dieses vorgibt, dass ein Wort zusammengeschrieben resp. getrennt werden sollte, dann machen wir das auch so auf der glossierten Ebene, und zwar unabhängig von der Schreibweise in der Originalnachricht.

Im Korpus gibt es sehr viele kreative Schreibweisen. An die Grenze dieser Möglichkeit geht ein Schreiber, der einen Zungenbrecher, der einen grossen Teil einer SMS ausmacht, ohne Leerschlag schreibt:

(39) *tschinchchatschedersvaunachatschadatschinchtschienttschinquauntatschinchchamuotschs*
("fünf Jäger jagen fünfhundertfündundfünfzig Gämsen")

Bei der Glossierung derartiger Konstruktionen treffen widersprüchliche Interessen aufeinander. Einerseits möchten wir die Konstruktionen der SMS-Schreiber möglichst nicht verändern. Andererseits hätten wir gerne Strukturen, die wir syntaktisch und lexikalisch analysieren können. Weiter möchten wir die glossierte Ebene möglichst nahe am kodifizierten Standard halten. Die erwähnten Konstruktionen entsprechen aber, auch wenn sie graphisch anders dargestellt werden, nicht dem Standard.

Neben diesen Überlegungen zur Handhabung des Korpus stellt sich auch die Frage nach der Intention des Schreibers. Wenn Buchstaben und Wörter ohne Leerschlag und Interpunktion aneinandergereiht werden, so stellt sich immer die Frage, ob dies aus Gründen der Ökonomie oder aus anderen, zum Beispiel stilistischen Gründen geschieht. Eine Schreibweise mit Bindestrich kann nie der Ökonomie geschuldet sein, denn das Schreiben eines Bindestrichs bedeutet immer mehr Aufwand als das Eingeben eines Leerschlages. Schlussendlich ist es auch so, dass Wortkomponenten, die durch Bindestriche getrennt werden, von RegEx (also der computerlinguistischen Syntax, die wir für die Suche im Korpus verwenden) als einzelne Wörter interpretiert werden, was bei Zusammenschreibung ohne Bindestrich nicht der Fall ist.
Aus allem Gesagten ergibt sich, dass folgende Regeln sinnvoll sind für das Zusammenschreiben bzw. Trennen von kreativen Wortkonstruktionen:

- Bei Konstruktionen, die mit Bindestrich getrennt sind, werden die einzelnen Komponenten in den Standard übertragen, dann aber wieder als Einheit zusammengesetzt – analog zu der Form, die in der Original-SMS auftritt.
- Konstruktionen, die zusammengeschrieben sind, werden nach den Regeln des Rumantsch Grischun getrennt und natürlich auch in die Standardformen übertragen.

18. Literaturverzeichnis

Darms, Georges (1994): “Zur Schaffung und Entwicklung der Standardschriftsprache Rumantsch Grischun”. In: Lüdi, Georges (Hrsg.) (1994): *Sprachstandardisierung*. 12. Kolloquium der Schweizerischen Akademie der Geistes- und Sozialwissenschaften 1991, Freiburg: Universitätsverlag: 3-21.

Decurtins, Alexi (2001) : *Niev vocabulari romontsch sursilvan – tudestg / Neues rätoromanisches Wörterbuch surselvisch-deutsch*. Chur: Societad Retorumantscha.

Eichenhofer, Wolfgang (2009): *Pledari sutsilvan – tudestg / tudestg – sutsilvan, Wörterbuch Sutsilvan – Deutsch / Deutsch – Sutsilvan*. Chur: Lehrmittelverlag des Kantons Graubünden.

Pledari Grond, Cuira: Lia Rumantscha, http://www.pledarigrond.ch

Schmid, Heinrich (1982): *Richtlinien für die Gestaltung einer gesamtbündnerischen Schriftsprache Rumantsch Grischun*. Chur: Lia Rumantscha.

Signorell, Faust (2001): *Vocabulari surmiran – tudestg / tudestg – surmiran, Wörterbuch Surmiran – Deutsch / Deutsch – Surmiran*. Chur: Lehrmittelverlag des Kantons Graubünden.

Tscharner, Gion (2007): *Dicziunari puter – tudas-ch / tudas-ch – puter, Wörterbuch Puter – Deutsch / Deutsch – puter*. Chur: Lehrmittelverlag des Kantons Graubünden.

Tscharner, Gion (2009): *Dicziunari vallader – tudais-ch / tudais-ch – vallader, Wörterbuch Vallader – Deutsch / Deutsch – Vallader*. Chur: Lehrmittelverlag des Kantons Graubünden.