University of Zurich[UZH]

URPP 'Language and Space'

Swiss German Dictionary – Part II

A research service for the Swisscom provided by the Language and Space Lab of the University of Zürich

Schmidt Larissa, Tandler Raphael, Mächler Alina, Samardžić Tanja

Contents

1 Outline of the Project

The project described here is the second phase of a research service provided from the University Research Priority Program (URPP) 'Language and Space' to the Swisscom Company. Contributors from the University (the UZH team) were the three students Raphael Tandler, Alina Mächler, and Larissa Schmidt. They were under the supervision of the Language and Space Lab director Tanja Samardžić. Correspondents from Swisscom were Claudiu Musat, Alexandros Lazaridis, Sandra Djambazovska, and Lucy Linder.

In the first part of the project, from October 2018 until January 2019, the Language and Space team created a Standard German to Swiss German dictionary. For 9000 Standard German words they generated Swiss German phonetical writings in six different Swiss dialects: Zürich, St. Gallen, Basel, Bern, Wallis, and Stans. The words were written in the Speech Assessment Methods Phonetic Alphabet (SAMPA). (From here onwards, a phonetic representation of a Swiss German word will be called a *SAMPA*.)

The second part of the research service, covered by the present report, proceeded from February 2019 until July 2019. The goal was to pair the Swiss German pronunciations from the first part of the project with Swiss German writings in the latin alphabet. (From here onward these written Swiss German words will be called *GSWs*.)

In the contract for the second phase it was agreed to add at least one GSW for the 9000 SAMPAs across the six dialects. It was additionally agreed to generate between 2000 and 3000 new SAMPAs for the same six dialects.

As a final result, we have now provided SAMPAs in six Swiss German dialects for 11248 Standard German words (9000 from the first phase, 2248 from the second). Concerning the writings, we manually wrote 9000 GSWs for Wallis and 2x9000 GSWs for Zurich and 600 for the four other dialects of St. Gallen, Basel, Bern, and Stans. The remaining writing variants are generated using automatic methods in agreement and collaboration with the Swisscom team.

In order to account for the multifarious ways in which Swiss German is written, we adapted our approach from generating manually one writing per SAMPA to instead deliver the means to automatically generate several writings per SAMPA. We tried to do this first by coming up with our own rule-based approach of a phoneme-to-grapheme script. The performance of this method was too low to be considered a satisfying project result. We thus engaged in experiments performed by the Swisscom team with the goal to train general sequence-to-sequence models on the pairs of SAMPAs and manually generated GSWs we had provided. The UZH team also provided evaluations for this model (see section 5 on page 8).

2 Goals

2.1 Problem: Variation in Swiss German writings

There are no orthographic rules for Swiss German writing. Therefore, the way people write Swiss German, varies. Even people that are speaking the same dialect, can write the same word differently.

For the GSWs in our dictionary, we considered choosing a way of writing that is widely used by dialectologists for Swiss German, namely the Dieth-writing set up by Eugen Dieth[1]. It focuses on one hand on phonetics and on the other hand on word-recognizability. This way of writing is intended primarily for linguists. But it deviates quite a lot from everyday Swiss German writing. It includes, for example, some specific diacritical signs that are certainly alien to non-linguists. Also, it does not represent the variety of writings used in everyday Swiss German writing.

2.2 Aim: Keeping the variation of Swiss German writings

For the GSWs in our dictionary, we decided for what we call "spontaneous" writings, i.e. writings as a Swiss German person might use them in everyday writing.

[1] See Dieth, Eugen (1986): *Schwyzertütschi Dialäktschrift: Dieth-Schreibung*. Serie Lebendige Mundart 1. Schmid- Cadalbert, Christian (ed.)

2.3 Solution

We came up with three approaches of how to generate Swiss German writings and keep the variation of them as broad as it is in use.

2.3.1 Approach 1: Manually

In the beginning, our UZH-team of native Swiss Germans generated spontaneous writings manually. Two sets of spontaneous manual GSWs are distinguishable:

- **Set 1**
  - 1x9000 GSWs for Wallis German (by Larissa)
  - 2x9000 GSWs for Zurich German (by Alina and Raphael)

These GSWs were written by native speakers of the respective dialect, and independently of the corresponding SAMPAs. By writing independently of SAMPAs we were able to avoid biases concerning the phonetics as well as the meaning of the word (for more on this see section 3.1 on page 5).

Our original goal was to have one GSW per all of the 9000 SAMPAs in each dialect. But, in order to account for the mentioned variation in writing, we wanted to have more than one spontaneous writing per SAMPA. Doing this all manually would have been to time consuming. We therefore opted to go for an automatic process that would take the SAMPAs from phase 1 together with the GSWs just described (set 1) as a start and generate a multitude of GSWs per SAMPA. Also, this automatic process should generate GSWs for the SAMPAs of the other four dialects for which we had not yet generated spontaneous writings, as we were not native speakers.

We first tried to automatise the generation of GSWs in a rule based way (see approach 2). Yet, recall was not satisfying with regards to set 1 (see 4.3.1 on page 6).

Finally a language model (approach 3 see 5 8) was set up by the Swisscom team. The model was trained on the GSWs from set 1 and the matching SAMPAs. The writings for the dialects of St. Gallen, Basel, Bern, and Stans which this model delivered were not good – mainly because there were some special SAMPA characters in those dialects and the model did not have the correlating latin character strings. To improve the model, we added the following manually written GSW.

- **Set 2**

For the annotation of the following dialects we had no native speakers. Therefore, when writing the GSWs, we relied on the SAMPAs in writing (see 3.1.2 on page 3.1.2).

- 600 GSWs for St. Gallen
- 600 GSWs for Basel
- 600 GSWs for Bern
- 600 GSWs for Stans

2.3.2 Approach 2: Automatically – rule-based

Our first attempt at generating GSWs automatically was to write a phoneme-to-grapheme script. This script took as input first the SAMPAs from project-part I. Second it took a list – which we wrote – of mappings between SAMPA characters and graphemic (namely latin) characters. The output of the script was a list of various different GSWs for each SAMPA. This first kind of automatic process was a rule-based approach, for which no training data was needed, but our knowledge contained in a set of written rules.

As already mentioned, the performance of this script was not satisfying with regards to set 1 (for more on this see 4.3 on page 6).

2.3.3 Approach 3: Automatically – model-based

The second and improved attempt to automatically get from one SAMPA to several GSWs, was Swisscom's setting up and training a character-level machine translation model. With the manual GSWs from set 1 and set 2 we delivered the required data for this model.

3 Spontaneous Swiss German writings (approach 1, used for approach 2 and 3)

Alina, Raphael, and Larissa manually wrote down some GSW in a spontaneous manner in order to check the output of approach 2 (set 1) and in order to train and improve the model of approach 3 (set 2) .

3.1 Writing "spontaneously"

The goal of writing GSWs manually was to do this in a spontaneous manner, i.e. how they would be written intuitively in a text message. We decided to only write one GSW per word, namely the first GSW that came to mind. We noticed that there always were a lot of versions that each of us deemed possible to write for one word. so we really had to restrict ourselves in not overthinking what the "best" GSW was. We came to the conclusion that there are several possible GSWs per SAMPA none of which could be deemed the best.

3.1.1 Set 1: Writing independently SAMPAs for Zurich and Wallis

The 2x9000 GSWs for Zurich and 1x9000 GSWs were done independently of the SAMPAs: Whilst writing them we only looked at the Standard German words, and not the respective SAMPAs.

Standard German Influence

Looking at the Standard German words whilst doing this might have tilted the way we wrote our spontaneous GSWs a bit to the Standard German side. Nevertheless the GSWs thusly generated are not implausible. Because from our personal experience we can say that a lot (though certainly not all) of Swiss people are influenced by the Standard German writing conventions in their Swiss German writing.

The scope of influence of Standard German writing on Swiss German writing might be of interest for further research. In our group we had different assumptions – one being that there is a high influence in less educated people since they learn to write Standard German in school and base most their written output on this knowledge. An opposing hypothesis was that there is a low influence in less educated and high influence for people at the university as they encounter Standard German more often.

Ambiguities

What also occurred to us was that the output of our phoneme-to-grapheme script would probably not match all of the spontaneous writings for Wallis and Zurich. For there were a lot of lexical and morphological ambiguities in the Standard German words, and we most likely did not always go for the same lexical or morphological (GSW-)options as we did in the first process of SAMPA-writing.

3.1.2 Set 2: Writing based on SAMPAs for St. Gallen, Basel, Bern, Stans

For these four dialects we wrote each 600 GSWs. As these dialects were not our native dialects we looked at the SAMPAs whilst writing the GSWs. So these writings were done in a less spontaneous manner. Still we chose the first writing option that came to our mind, but we were somewhat predisposed.

- **Bias concerning the phonetics of the word**: For example, for the Standard German 'natürlich', when we saw the SAMPA / t y: R l I /, we wrote the GSW *türli*. Not having seen SAMPA, we might have chosen *natürli* which is the longer form of the same word.
- **Bias concerning the meaning of the word**: The Standard German 'liebe' is morphologically ambiguous. It can (as we did not differentiate upper- and lowercase) both mean (a) '(I) love' or (b) '(the) love)'. Seeing either the SAMPA (a) / l I{ b I / or (b) / l I{ b I / did predispose us concerning whether we chose the GSW (a) *liäbi* or (b) *liäbe*.

4 From phonetical representations of Swiss German words to their writings. A rule-based program (Approach 2)

What follows is a description of our approach 2 to the automatic generation of GSWs. Via SAMPAs together with phoneme-to-grapheme mappings we tried to yield possible GSWs.

4.1 Broad and narrow rules (BR and NR)

We created a list of mappings connecting SAMPA characters with latin characters. In what we called the "broad rules" (BR) we wrote down every mapping that we deemed possible (see Appendix 2). So for example, SAMPA /tS/ could go with the latin characters *tsch,ts,tz,* or *tch*. For the narrow rules we chose mostly only one graphemic option (max. 4) that seemed the most straightforward to us – in this example only *tsch* (see Appendix 1 for the narrow rules).

4.2 Script phoneme-to-grapheme

We then wrote a Python script that took as input a list of SAMPAs together with either BR or NR. The output was a list of (what should be) possible writings (GSWs) for each phonetical representation (SAMPA) of a Swiss German word.

4.3 Evaluation of Swiss German writings from rule-based program

Naturally the scripts output, when using BR, was bigger. For the first 2828x6 SAMPAs the script gave out:

- using BR: roughly 28 mio. GSWs
- using NR: roughly 1 mio. GSWs

We evaluated parts of that output. Our evaluation is presented below.

4.3.1 Recall

We used set 1 of our manually generated spontaneous writings (mentioned in 3.1.1) to evaluate the recall of our phoneme-to-grapheme script. Set 1 was written before the script. And, as already mentioned, set 1 was done independently of SAMPA. This was important in order to allow these spontaneous GSWs to be a means of testing the recall of the automatically generated GSWs.

4.3.2 Error analysis and correction

We evaluated the recall of the output of our phoneme-to-grapheme-script against the before-mentioned spontaneously written GSWs. We checked 2828 out of 9000 words.

- Of Alina's 2828 spontaneously written ZH words
  - for BR: 2180/2828 were matched
  - for NR: 1767/2828 were matched
- Of Raphael's 2828 spontaneously written ZH words
  - for BR: 2310/2828 were matched
  - for NR: 1918/2828 were matched
- Of Larissa's 2828 spontaneously written VS words
  - for BR: 2397/2828 were matched
  - for NR: 2015/2828 were matched

We had expected that recall would be lower when using NR than BR, as NR often only contained one latin character string per SAMPA character. And indeed, recall was around 2/3 using NR, and higher, around 4/5, using BR. For BR, our aim was to create an exhaustive list of latin characters for every SAMPA character. For that reason, we had expected that recall would be 100 percent when using BR. This turne out not to be true.

We had a detailed look at some of the spontaneous words (from set 1) that were not matched by the script's output (using BR) and noted down possible causes for this:

- **Issues with SAMPA writing**
  We noticed that there are some mistakes in the SAMPAs, that could still be corrected. Correcting all these mistakes would have been way too time consuming for our purposes of quick resolutions to improve recall. Some of the mistakes in SAMPA were are an easy fix though:
  - **Lack of spaces in GSWs**
    For some GSWs, people would conventionally add a space, for example when there are two different words: The Standard German preterite corresponds to the Swiss German past participle – 'ging' would (most often) be written *isch gange* in spontaneous writing.
    Yet, the phonetic representation of the above example is / I S k a N @ /. Applying transliteration rules thus gives *ischgange*. To facilitate further mappings to writing, we introduced explicit boundary making: For the new SAMPAs of phase 2, we inserted an _ in the SAMPA where there should be a space in writing. In this example, the SAMPA would be / I S _ k a N @ / .
  - We had a lot of /t s/ in the SAMPAs from the first

delivery and noticed that these actually most often should have been /ts/. So, we corrected all the / t s / to / ts / (even though in rare instances, i.e. at word-borders where *t* and *s* meet it would really be / t s /, for example in *post-sache*).

- **Slips of the pen within spontaneous words**.
  Deemed not useful to correct, and therefore not done.
- **Ambiguities of Standard German Word**.
  We sometimes decided for different options in SAMPA and in spontaneous writing. This is so because, first, we did our SAMPAs and our spontaneous GSWs (at least set 1, see sec. 2.3.1) independently of each other. And secondly, due to lexical or morphological ambiguities there are for the same Standard German word often several Swiss German options.
  What could be done: In the first part of the project, we marked ambiguous words. For every word that is ambiguous, all their versions could be introduced. With the consent of the Swisscom team, this is left for future work.
- **Variation in the dialect**.
  - For example short vs. long versions as in *türli* and *natürli* which both translate to Standard German 'natürlich'.

  Again, we sometimes chose different versions when doing SAMPA or spontaneously writing. Introducing all the different versions to the SAMPAs would have been too time consuming, therefore not done.
- **Mappings that are missing in BR**:
  We thought about introducing more mappings to BR, which we might have missed. For example, we thought about adding a correspondence from SAMPA /v/ to graphemic *v* (as it occurs in SAMPA / a kx t i v i t {: t @ / to GSW *aktivitäte*). Introducing even more mappings, and therewith way more GSWs that were wrong in most cases, was deemed not a good idea and we did not do this.
- **Pronunciation vs. writing**:
  In some cases, Swiss German writing did not map to the pronunciation as expected, i.e. according to our NR or BR. There might be various causes for this:
  - Standard German influence: A person from Zurich says / g o: s aU / (with a lenghtened vowel and short consonant s) but writes, as in Standard German *Gossau* (with double s).
  - foreign words, f.ex. french or english words, might follow the foreign orthography which we did not take into account in our assignation rules.

4.3.3 Precision

One option we considered for evaluating the GSWs generated from automation was to (automatically) check for the occurrence of each GSW in existing corpora (text collections) of Swiss German writings. Here are some examples of Swiss German writing corpora:

- *sms4science.ch*
  (SMS Collection of the University of Zurich)
- *https://www.chtk.ch/index.php/de/*
  (collection of older texts)
- *https://www.idiotikon.ch/literatur/dialektwoerterbuecher*
  (Lexica of individual dialects)
- *http://textberg.ch/site/de/willkommen/*
  (SAC yearbooks)

We decided not go for corpora and instead evaluating some instances manually as getting authorizations for corporate usage and reformatting the corpora to fit our needs would have taken a lot of time.

The precision was very low. This should be visible from Appendix 3. We went through all the GSWs which the script wrote out for the first thirty words and most of them seemed impossible to us (marked red in App. 3). Only very few seemed like good writings (marked green), and some seemed ok, but rather unusual (marked orange).

As an improvement for precision, we considered limiting the mappings of BR and NR. We wanted to go from character-level mappings to cluster-level mappings. We had already begun seeking out some often occurring consonant clusters.

But then (in the meeting on the 7th of May) we realized that everyone's time could be used in a more efficient way: We agreed that the Swisscom team proceed on improving the automation as they are the experts in that field, and Alina's, Raphael's, and Larissa's time should rather be used within their dialectal expertise, namely in evaluating GSW output of Swisscoms automation, and in generating more SAMPA.

5 Evaluation of Swisscom's Machine learning (approach 3)

At a meeting of the 7th of May Claudiu presented an alternative to our rule-based approach: The possibility of training a character-level machine translation model on the existing mappings of SAMPA and GSW. The output of this model should be a list of n-best writings for every SAMPA string. The machine-learning should have the advantage of generally producing better output; especially since it could potentially capture the exceptions to the general rules that were applied in the rule-based approach.

On the side of Swisscom, it was Lucy Linder and Sandra Djambazovska who were responsible for the development of the translation model.

Also at the meeting of the 7th of May, it was decided that the output of the machine model will then be evaluated by the UZH group and it would be manually checked if the proposed candidates are plausible. For the evaluation, our team formulated some criteria in order to consistently rate the output. Generally, the criteria aimed to exclude only very far-off forms and accept such that contain one or two misspellings.

5.1 Guidelines for GSW-evaluation

The guidelines are the following: A GSW is wrong (tagged "0") when there's one typo or two orthographic mistakes. Otherwise, it is deemed correct (tagged "1"). (Mind you that GSWs tagged with 1 are not necessarily "good" writings, they might include some writings that seem off to the Swiss German reader. Our goal was mainly to exclude those writings that are very probable to be seen as false.)

Typo:

- Addition of a single letter without comprehensible reason e.g. *pfanep*, *pcer*, *pflanzle*
- Missing a letter without comprehensible reason and not due to phonetic contractions *pter* vs. *vezelle*/*vrzelle*
- Changing of a letter without comprehensible reason e.g. *ensionskasso*,*rosknacht* vs. *rosknächt*, *rosknecht*
- Special cases of sounds that are related but still count as a typo: *ch* instead of *ck*, e.g. *wipchingu*, *c* for *s*

Orthographic mistakes:

- Exchange of related sounds/spellings: *b-p*, *t-d*, *g-k*, *v-f*,*v-w*
- Added or omitted geminates: *penssionskase*
- Changing of vowel length (prolonging/shortening): *peeter*, *kle*

5.2 Test-set and evaluation

Our test set was 1000 of the 9000 words of phase 1. What we had delivered in that phase was split into for lists, list 1, 2, 3.1, and 3.2. All of those lists themselves were ordered alphabetically. The last 1000 words of list 3.2 (from Standard German 'pensionskasse' to 'zwischenzeitlich') were the test-set.

For each of these 1000 words in each of the six dialects, we received 5 predictions for GSWs from the Swisscom Group. We evaluated the following number of GSWs but did not yet do any counting of these 0/1 tags:

- For Wallis and Zurich we tagged each 1000x5 GSW predictions with 1 or 0.
- For St. Gallen, Basel, Bern, and Stans, we did each 200x5.

6 New SAMPA

As mentioned in the outline of the project, the goal was to produce a number of additional SAMPA writings for the six dialects. A list of 3431 words was provided by Swisscom. Some words had to be deleted since there were doubles within the list, doubles across lists or nonsensical words. Finally, 2248 words were done for all six dialects.

7 Concluding Remarks

The initial agreement of the Language and Space Lab team and Swisscom was to deliver one writing for the 9000 words across the six dialects, as well as 2000 to 3000 additional SAMPAs. The SAMPAs were delivered as agreed.

- Additionally to the 9000x6 SAMPAs of project-part I, we wrote 2248 new SAMPAs for the six dialects of Zurich, Basel, Bern, Stans, St. Gallen, and Wallis each.

Concerning the production of GSWs the Language and Space Lab team did not follow exactly the approach as it was fixed in the contract. Instead, we only delivered some manually generated GSWs, namely

- for the 9000 SAMPAs of Zurich we delivered two versions of manually generated GSWs (spontaneous writings of Alina and Raphael)
- for the 9000 SAMPAs of Wallis, we delivered 9000 manually written GSWs (spontaneous writings from Larissa
- To improve Swisscoms language model, we additionally delivered 600 GSWs for the other dialects (but, as the dialects of St. Gallen, Basel, Bern, and Stans are not our native dialects, we did this on the base of the SAMPAs and could not do it independently).

Also, we provided some evaluations for the predicted GSW that the Swisscom group gave us based on their model:

- For Wallis and Zurich we tagged each 1000x5 GSW predictions with 1 or 0.
- For St. Gallen, Basel, Bern, and Stans, we did each 200x5.

With all this we delivered the means to gain a lot more than one possible GSW per SAMPA. Our team ended up doing less manually written writings than agreed but provided the base to generate more correct possibilities through a model trained program.

Appendix 1 – NR

tS	tsch
y:	ü
O:	oo
oU:	ou
ll	ll
ts	ts,z,tz
2:	ö
th	th,t
@	e,ä
O~	o,on
y{	üä
Ou	ou
2I	öi
d	d
OI	oi,eu
h	h
yE	üe
l	l,ll
p	p,b,pp
t	t,d,tt
y@	üe
x	ch
OU	ou
ei	ei
U:	uu
{U:	äu
eU:	eu
eI	ei
rr	rr
AI	ai,ei
o:	o
E:	e,eh,ä,äh
eU	eu
C	ch
EI	ei
oI	oi,eu
O	o
S	sch
IU:	iu
e:	e,ee,eh
EU	äu
oU	ou
Ei	ei
g	g
k	k,g,gg
o	o
o~	o,on
s	s,ss
E~	e,en
UI	ui
Ai	ai,ei
w	u,l,ll
UE	uä
AU	au

{	ä,e
tt	tt
9i	öi
pS	psch,bsch
Ue	ue
xx	ch
9I	öi
pp	pp
u:	u
2	ö
pf	pf
Y@	üe
ph	p,ph,bh
AU:	au
EU:	eu
N	ng,n
9:	ö
uI	ui
u@	ue
b	b
oi	oi,eu
f	f,ff,v
mm	mm
j	j,i
n	n,nn
ss	ss
r	r,rr
v	w
ue	ue
ua	ua
i{	iä, ie
{:	ä,äh,e,eh
au	au
io	io
a~	a,an
o~:	o,on
I:	i,ie
nn	nn
aI	ai,ei
iU	iu
Y{	üä
Y:	ü
aU	au
9	ö
i@	ie
A:	a
iE	ie
A	a
SS	sch
i:	i,ie
E	e,ä
{u	äu
I	i
IU	iu

{~	ä,än,e,en
kS	gsch
Eu	eu,äu
U	u
ff	ff
Y	ü
I@	ie
a:	a
IE	ie,iä
ou	ou
a	a
I{	iä
e	e
{U	äu
i	i
kh	k,g,gg
m	m,mm
U@	ue
ks	gs,x
R	r
u	u
Io	io
y	ü
{I	äi,ei
kx	k,ck
A~	a,an

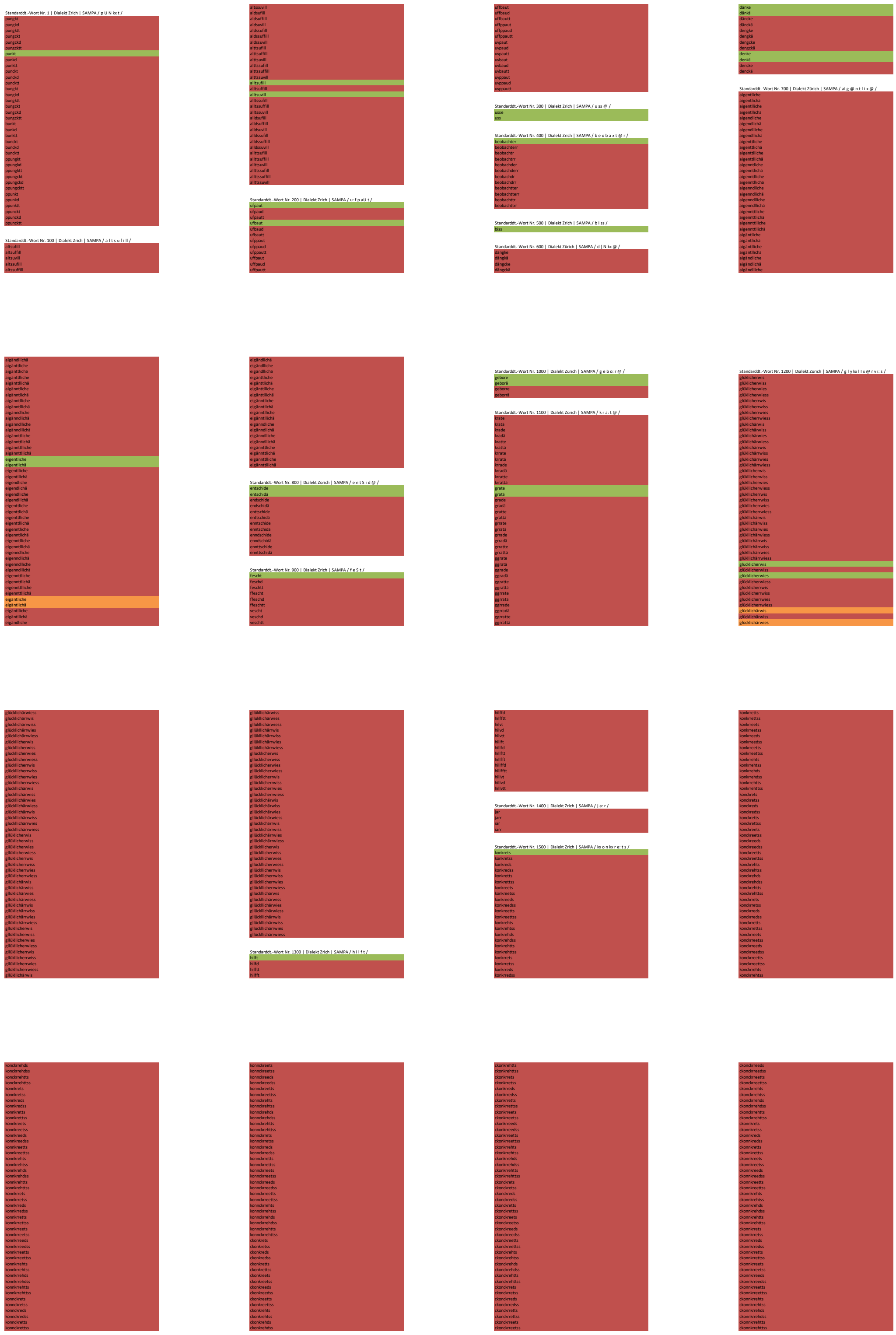
Appendix 2 – BR

```
tS      tsch,ts,tz,tch
y:      üü,ü,üh
O:      oo,o,oh
oU:     ou,ouu, ouh
ll      ll,l
ts      ts,z,tz
2:      ö,öö,öh
th      th,t,d
@       e,ä
O~      o,on,ong
y{      üä,üe
Ou      ou
2I      öi
d       d,t
OI      oi,eu
h       h
yE      üe,üä
l       l,ll
p       p,b,pp,bb
t       t,d,tt
y@      üe,üä
x       ch,k,ck
OU      ou
ei      ei,ai
U:      u,uu,uh
{U:     äu,äuu,äuh,eu,euu,euh
eU:     eu,euu,euh
eI      ei,ai
rr      r,rr
AI      ai,ei
o:      o,oo,oh
E:      e,ee,eh,ä,ää,äh
eU      eu,oi
C       ch,k,ck
EI      ei,ai
oI      oi,eu
O       o
S       sch,sh,s
IU:     iu,iuu,iuh
e:      e,ee,eh
EU      eu,äu
oU      ou
Ei      ei,ai
g       g
k       k,gg,g,ch,ck
o       o
o~      o,on,ong
s       s
E~      e,en,eng
UI      ui
Ai      ai,ei
w       w,u,l,ll
UE      ue,uä
AU      au
```

```
{       ä,e
tt      t,tt
9i      öi
pS      psch,psh,bsch,bsh
Ue      ue
xx      ch,ck,chch
9I      öi
pp      p,pp
u:      u,uu,uh
2       ö
pf      pf
Y@      üe,üä,öe,öä
ph      p,ph,bh
AU:     au,auu,auh
EU:     eu,euu,euh
N       ng,n
9:      ö,öö,öh
uI      ui
u@      ue,uä
b       b,p,bb,pp
oi      oi,eu
f       f,ff,v
mm      m,mm
j       j,i,y
n       n,nn
ss      s,ss
r       r,rr
v       w
ue      ue
ua      ua
i{      iä,ie
{:      ä,ää,äh,e,ee,eh
au      au
io      io
a~      a,an,ang
o~:     o,on,ong
I:      i,ii,ih,ie
nn      n,nn
aI      ai,ei
iU      iu
Y{      üä,üe
Y:      ü,üü,üh
aU      au
9       ö
i@      ie,je,iä,jä
A:      a,aa,ah
iE      ie,iä,je,jä
A       a
SS      sch,schsch,sh,shsh
i:      i,ii,ih,ie
E       e,ä
{u      äu,eu
I       i,ie,ii
IU      iu,ju
```

```
{~      ä,än,äng,e,en,eng
kS      gsch,ksch
Eu      eu,äu
U       u
ff      f,ff
Y       ü,ö
I@      ie,iä
a:      a,aa,ah
IE      ie,je,iä,jä
ou      ou
a       a
I{      iä,jä,ie,je
e       e
{U      äu,eu
i       i,ii,ie
kh      kh,gh,k,gg
m       m,mm
U@      ue,uä
ks      gs,ks,x,xs
R       r
u       u
Io      io,jo
y       ü
{I      äi,ei
kx      kch,gch,k,ck
A~      a,an,ang
```

Appendix 3 – Precision of Output for first 30 words



Appendix 3 – Precision of Output for first 30 words