swiss german NLP

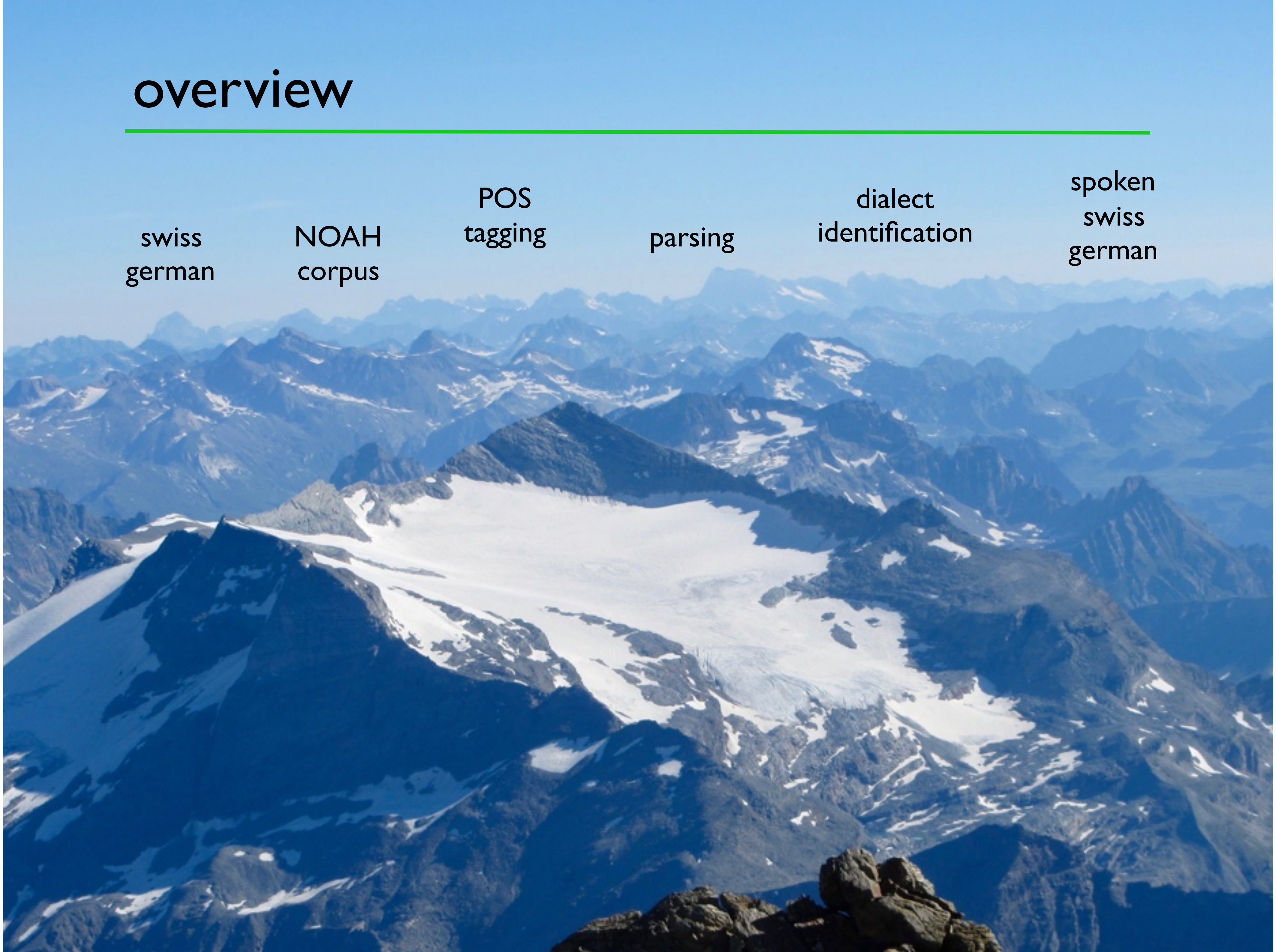
Nora Hollenstein & Noëmi Aepli

nho@zurich.ibm.com

noemi.aepli@uzh.ch

IBM Developer UnConference 18.01.2018

overview



swiss german

NOAH corpus

POS tagging

parsing

dialect identification

spoken swiss german

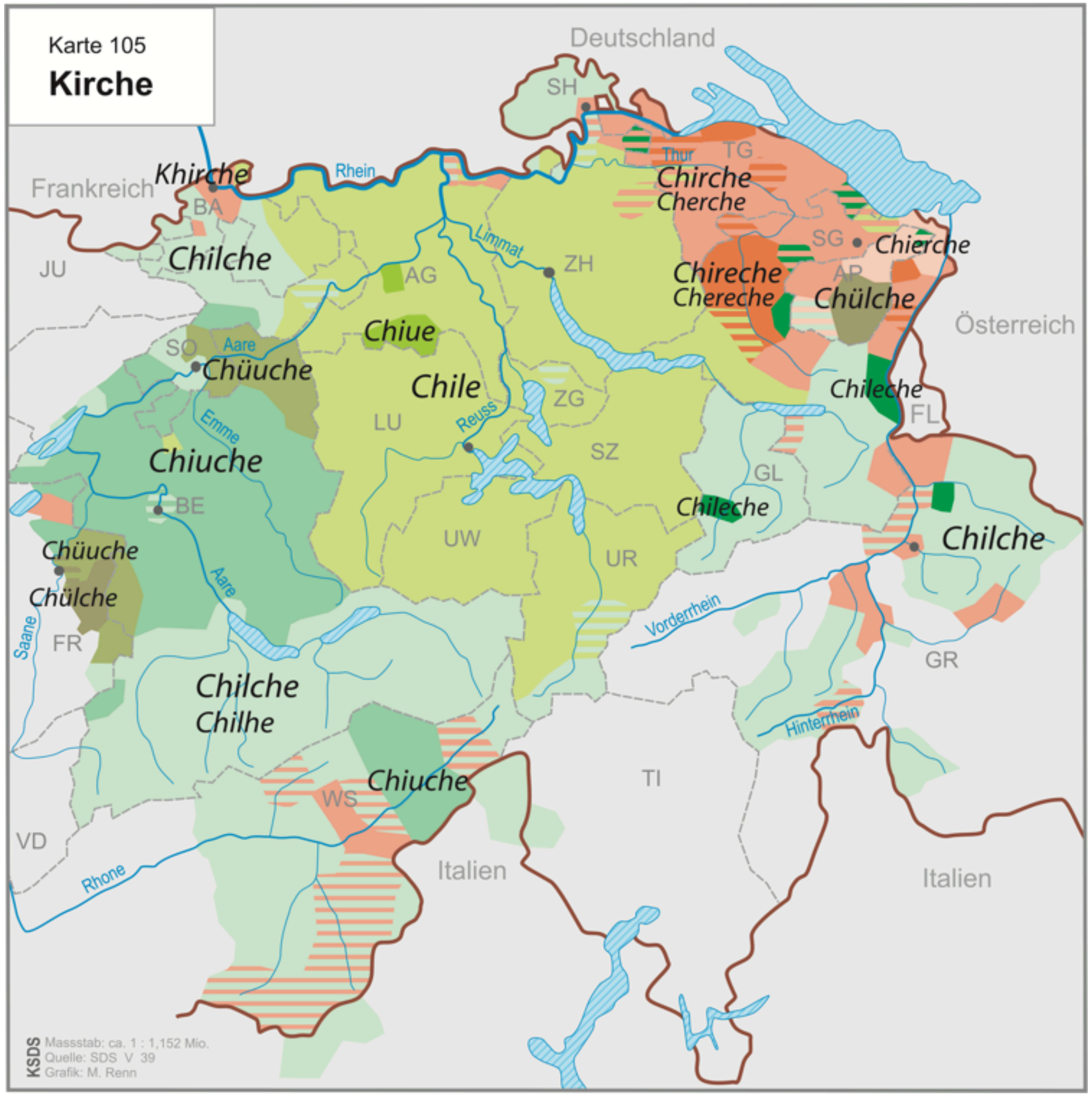
swiss german

- differences in every (linguistic) aspect
- dialects vs. standard german
- dialect vs. dialect

K
Kater
Pferd
Spinat
Kuss
Ross
K
B
Butter
K
Naastüechli
Zibele
A
Aprikose
Schmetterling
A
R
Rööschti
Gänseblüemli
R
G
G
Gurke
G

source: https://www.nzz.ch/nzzas/nzz-am-sonntag/richtig-krass-diese-sprache-ld.2615

swiss german



lexical &
morphological
differences

GSW vs. **GSW**

en/es kafi

GSW vs. **DE**

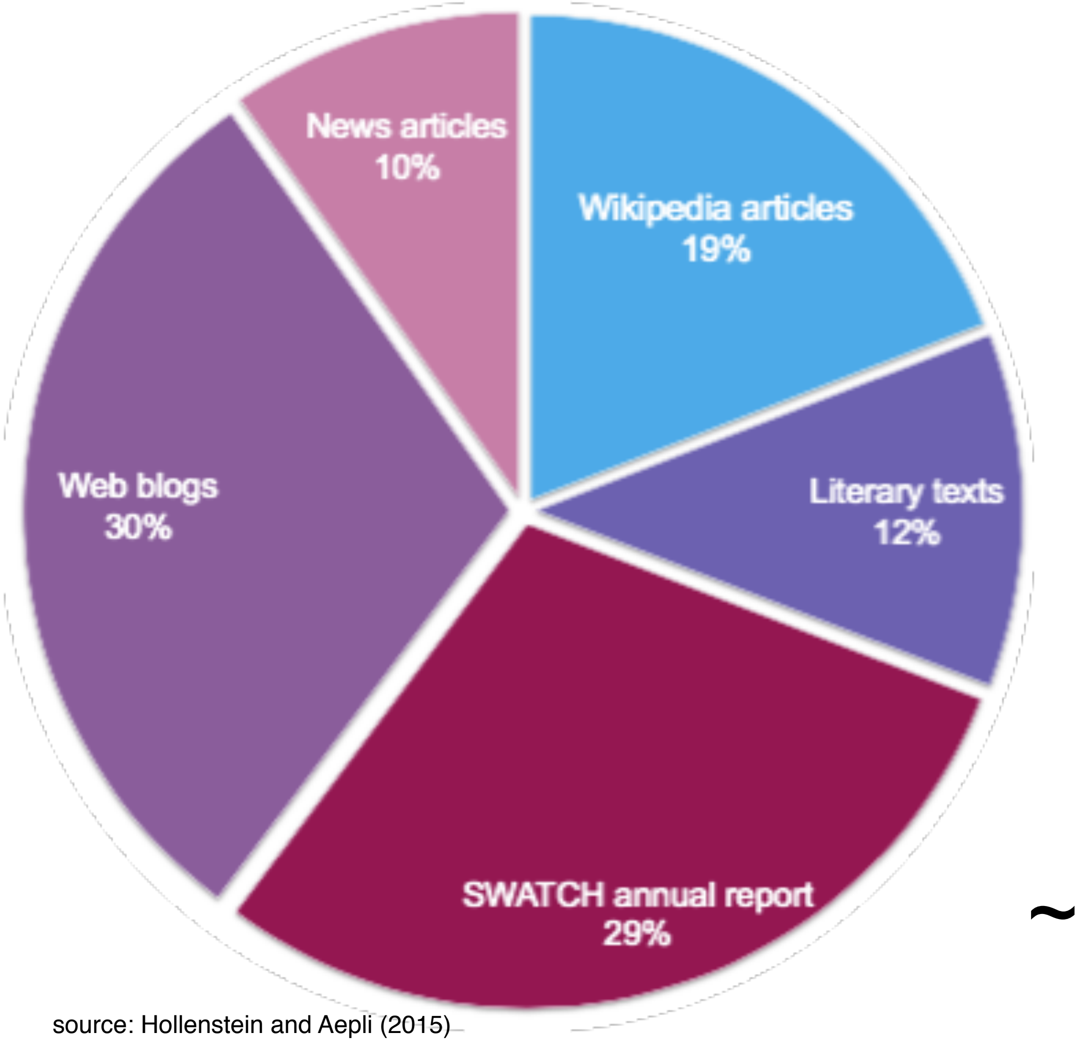
die schnecke

de schnägg

source: Kleiner Spachatlas der deutschen Schweiz: http://www.ksds.uzh.ch/de.html

NOAH corpus

... of written GSW



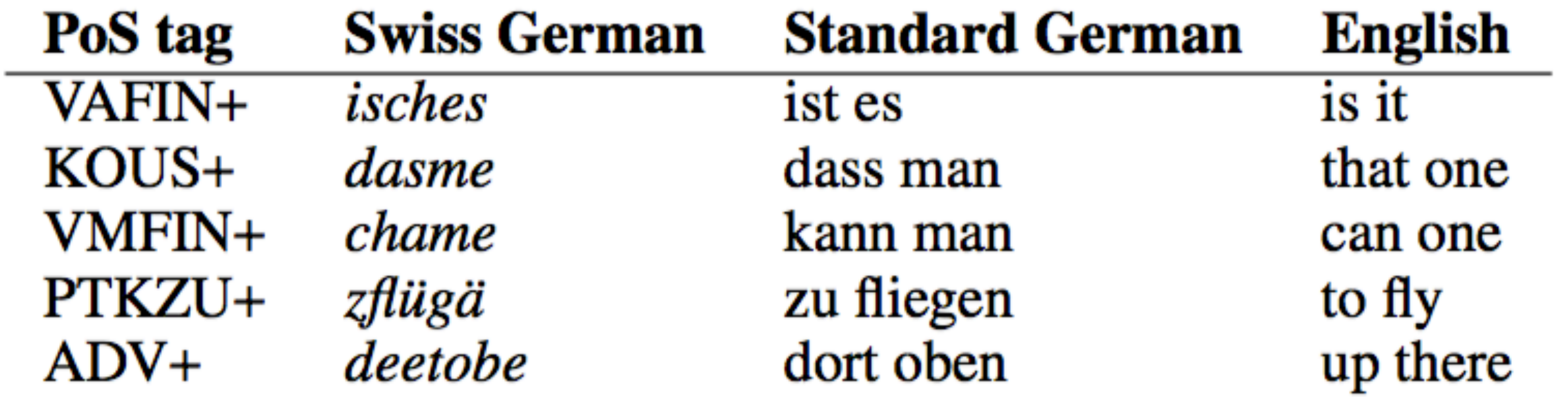
source: Hollenstein and Aepli (2015)

~116'000 tokens

POS annotated

POS tagging

- **STTS** 54 part-of-speech tags for standard DE
- **PTKINF** ich gòò ez **go** pòschte
  source: Glaser (2003)
- **TAG+**



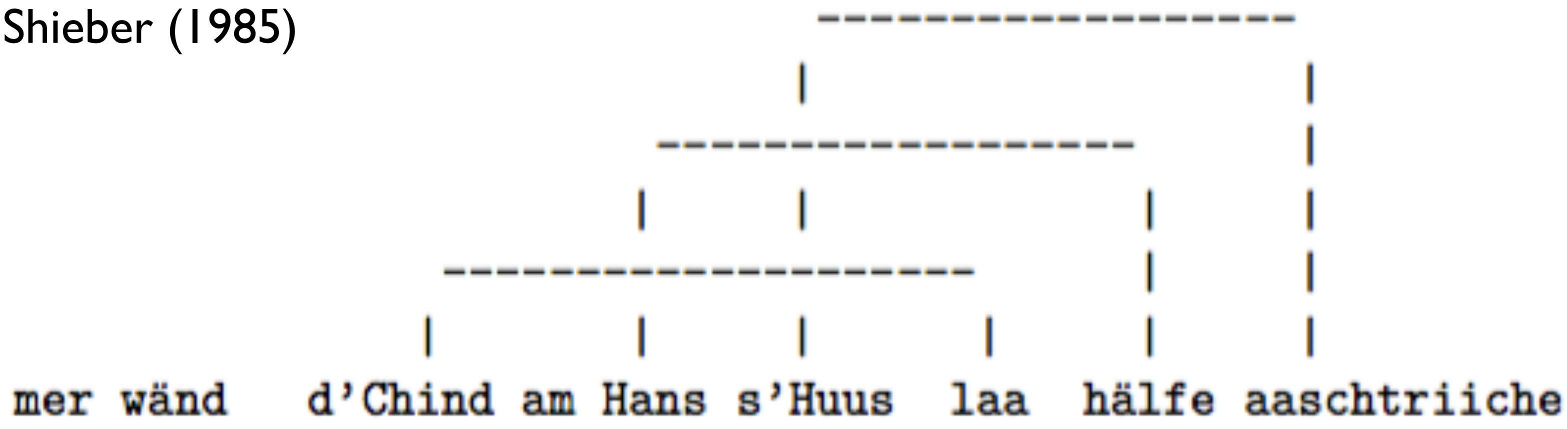
PoS tag	Swiss German	Standard German	English
VAFIN+	*isches*	ist es	is it
KOUS+	*dasme*	dass man	that one
VMFIN+	*chame*	kann man	can one
PTKZU+	*zflügä*	zu fliegen	to fly
ADV+	*deetobe*	dort oben	up there

source: Hollenstein and Aepli (2014)

parsing

a context-sensitive language (?)

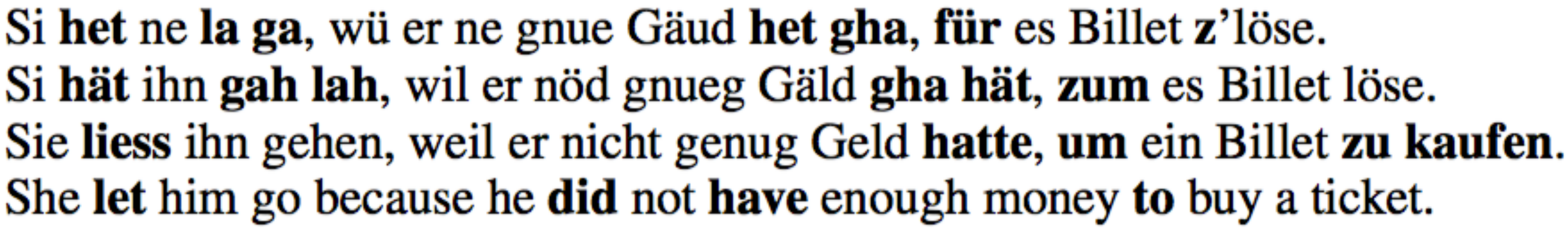
Shieber (1985)



mer wänd d'Chind am Hans s'Huus laa hälfe aaschtriiche

source: https://files.ifi.uzh.ch/cl/siclemat/lehre/hs09/ecl1/script/script.pdf

parsing



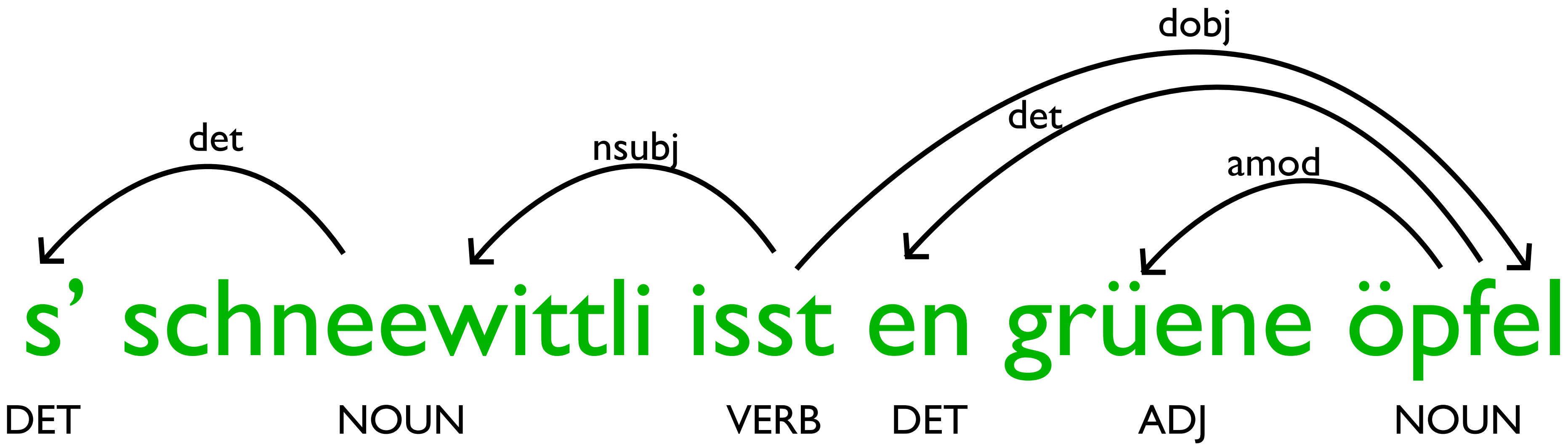
“BE” Si **het** ne **la ga**, wü er ne gnue Gäud **het gha, für** es Billet **z’**löse.
“ZH” Si **hät** ihn **gah lah**, wil er nöd gnueg Gäld **gha hät, zum** es Billet löse.
DE Sie **liess** ihn gehen, weil er nicht genug Geld **hatte, um** ein Billet **zu kaufen**.
EN She **let** him go because he **did** not **have** enough money **to** buy a ticket.

source: Hollenstein and Aepli (2014)

syntactic differences

- word ordering
- final clauses
- tenses
- cases
- overt subj.
- ...

parsing



s' schneewittli isst en grüene öpfel

DET NOUN VERB DET ADJ NOUN

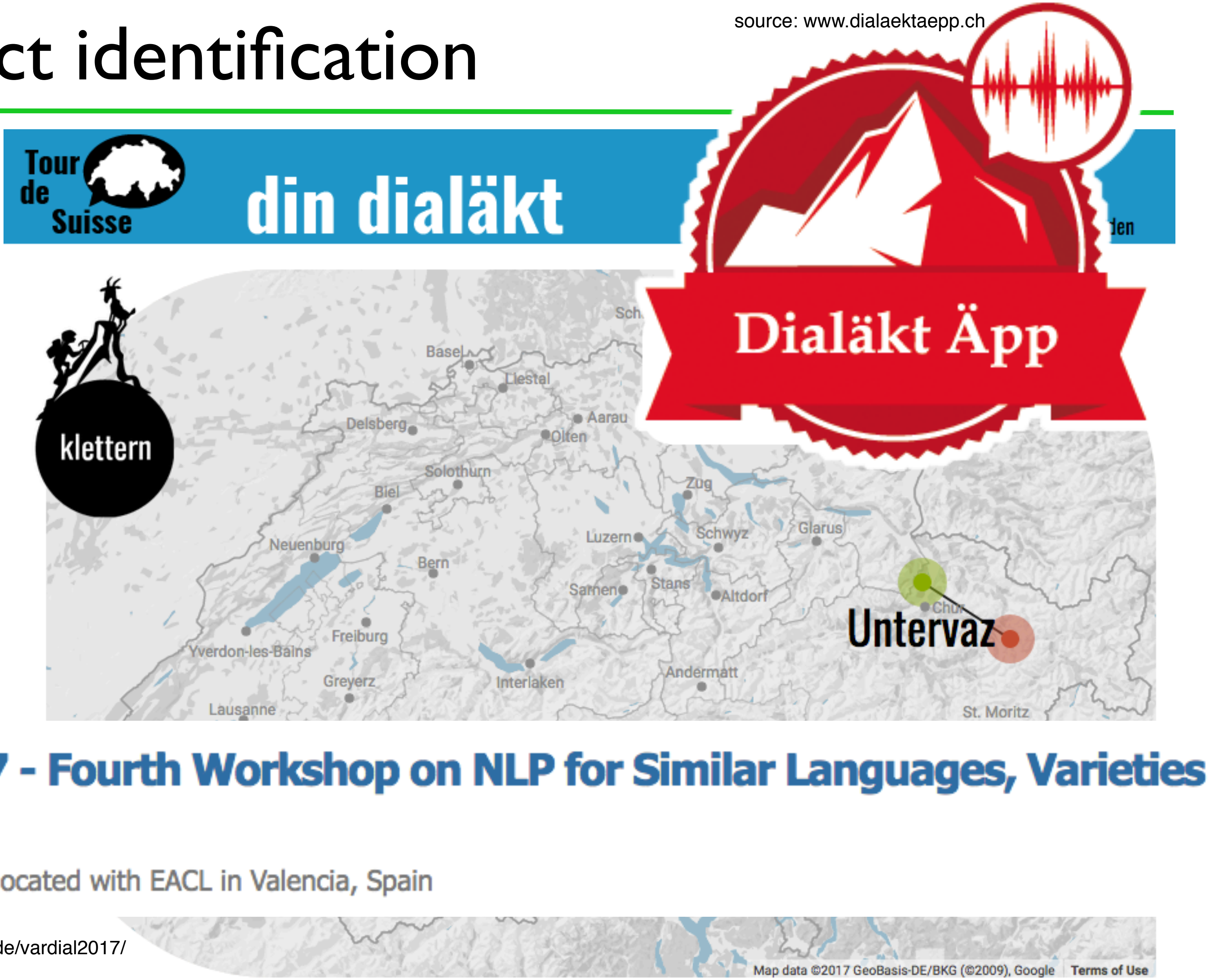
goal universal dependencies for swiss german

approaches annotation projection & model transfer

result ~60% LAS on 100 gold standard sentences

dialect identification

source: www.dialaektaepp.ch



VarDial 2017 - Fourth Workshop on NLP for Similar Languages, Varieties and Dialects

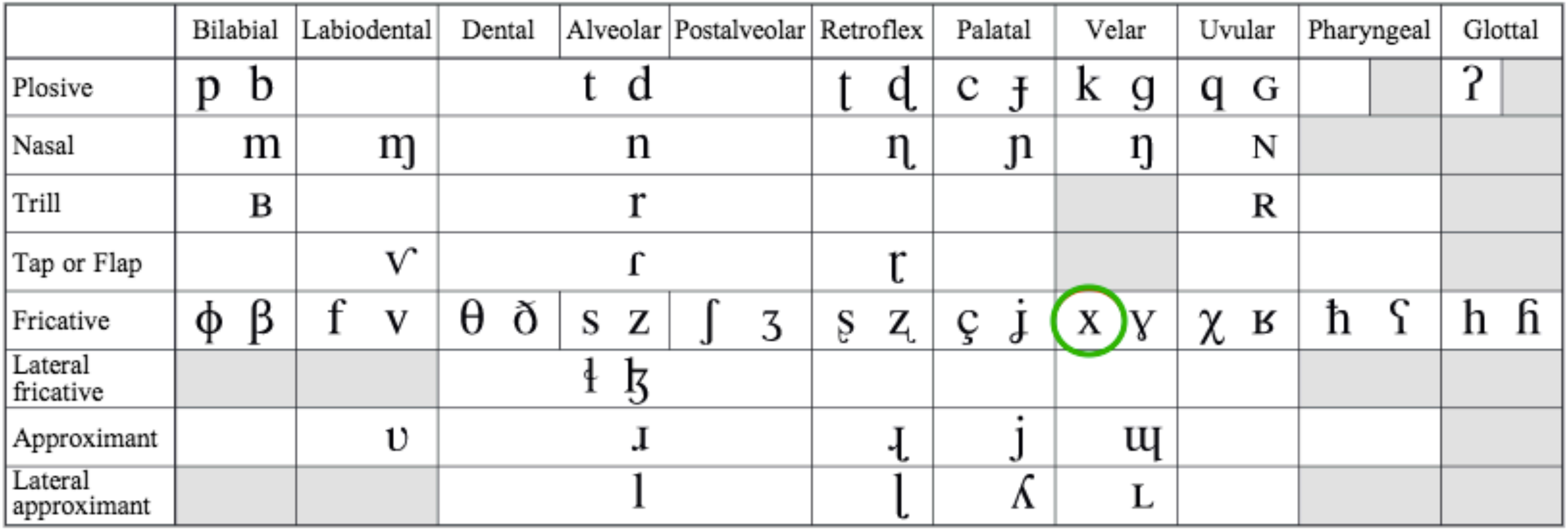
April 3rd, 2017 - Co-located with EACL in Valencia, Spain

source: http://ttg.uni-saarland.de/vardial2017/

source: www.dindialaekt.ch

spoken swiss german

GSW vs. DE

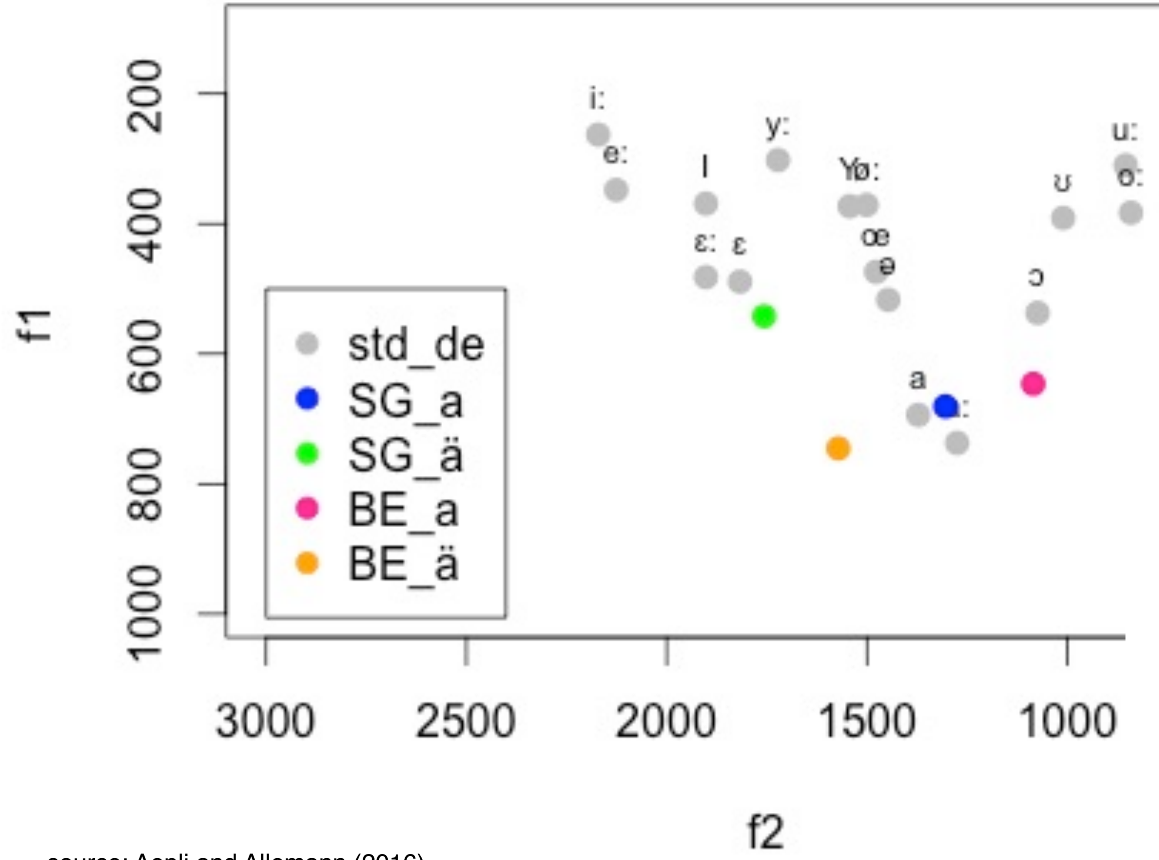


	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

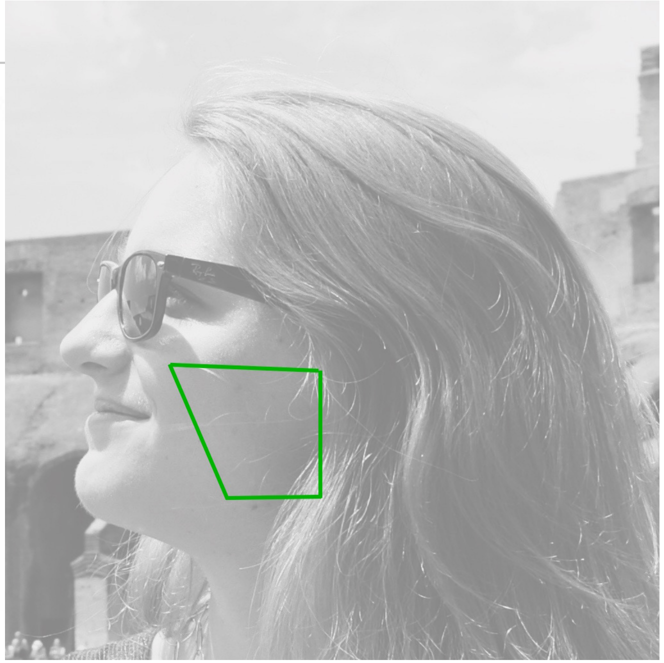
source: https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf

spoken swiss german

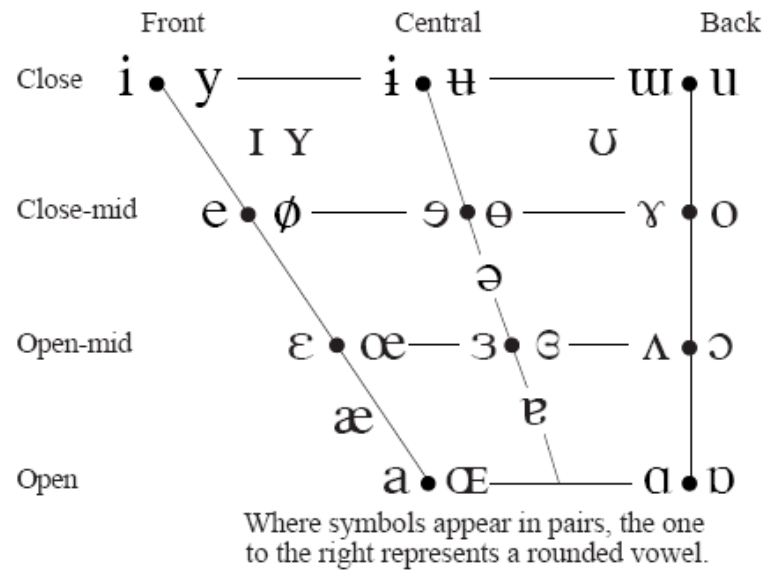
GSW vs. GSW



source: Aepli and Allemann (2016)



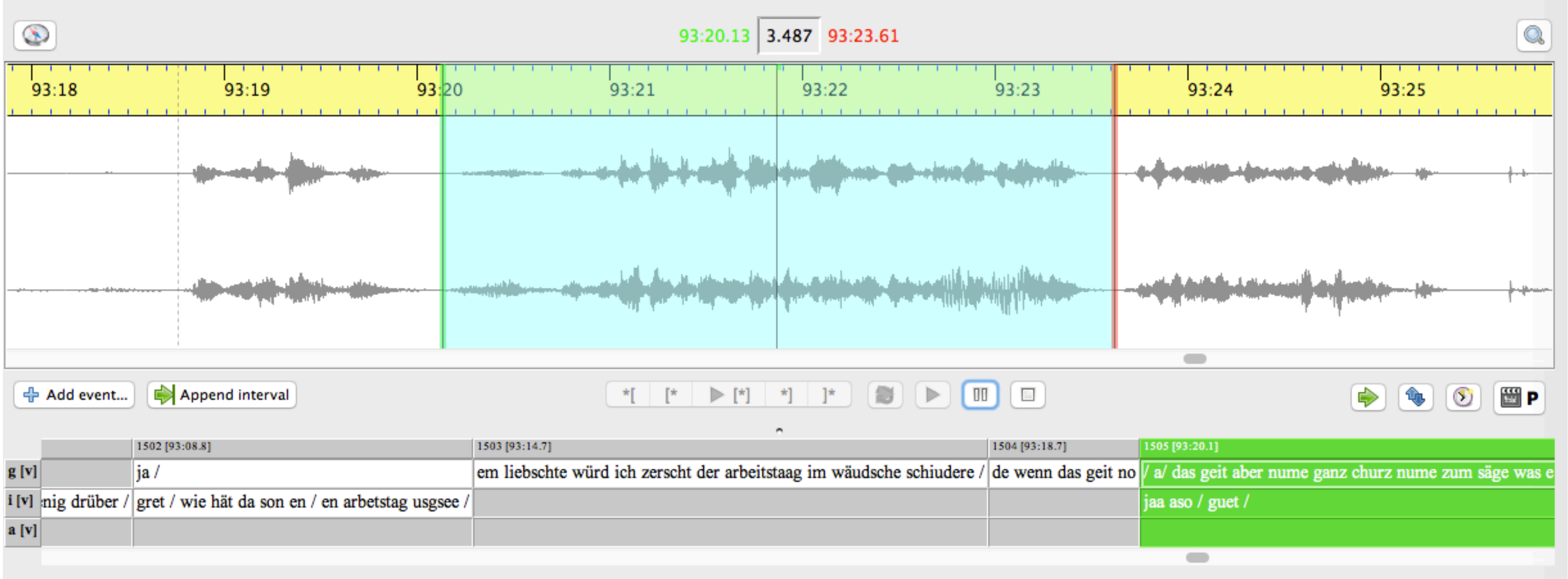
VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

source: https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf

spoken swiss german



		1502 [93:08.8]	1503 [93:14.7]	1504 [93:18.7]	1505 [93:20.1]
g [v]		ja /	em liebschte würd ich zerscht der arbeitstaag im wäudsche schiudere /	de wenn das geit no	/ a/ das geit aber nume ganz churz nume zum säge was e
i [v]	nig drüber /	gret / wie hät da son en / en arbetstag usgsee /			jaa aso / guet /
a [v]					

screenshot: transcription tool EXMARaLDA

ArchiMob corpus

53 transcribed videos, POS annotated, normalised

conclusions

- compilation of resources for GSW **dialect research**
- development of basic NLP tools for dialect research
- approaches **generalisable** to lower resourced languages
- **applications** in industry → conquer swiss market ;)

resources

- NOAH Corpus https://gitlab.cl.uzh.ch/noah/corpus
- ArchiMob Corpus http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html
- dindialaekt.ch https://www.dindialaekt.ch/tour-de-suisse/de

  ⟶ GSW - DE translation ("aufschreiben" > DE)
- VarDial 2017 http://ttg.uni-saarland.de/vardial2017/
- dialäkt äpp http://www.dialaektaepp.ch/
- ... for more, check out: https://gitlab.cl.uzh.ch/noah/corpus

literature

Hollenstein, N. and Aepli, N. (2014). "Compilation of a Swiss German Dialect Corpus and its Application to PoS Tagging". *International Conference on Computational Linguistics (COLING 2014)*, page 85.

Hollenstein, N. and Aepli, N. (2015). "A Resource for Natural Language Processing of Swiss German Dialects". In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL 2015).*

Aepli, N. and Allemann, A. (2016). "Schwiizer{d|t}ütschi Vokä{u|l} – west vs. ost". Seminar Thesis.

Samardžić, T., Y. Scherrer, E. Glaser (2016). "ArchiMob - A Corpus of Spoken Swiss German". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Portorož, Slovenia.

Samardžić, T., Y. Scherrer, E. Glaser (2015). "Normalising Orthographic and Dialectal Variants for the Automatic Processing of Swiss German", In *Proceedings of the 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznan, Poland.

Zampieri, M., Malmasi, S., Ljubešic´, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). "Findings of the Vardial Evaluation Campaign 2017". In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial),* pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Glaser, E. (2003). "Schweizerdeutsche Syntax: Phänomene und Entwicklungen". In Dittli, Beat; Häcki Buhofer, Annelies & Haas, Walter (Hrsg.): "Gömmer MiGro?" Freiburg, Schweiz, 39–66.

Shieber, S. M. (1985). "Evidence Against the Context-freeness of Natural Language". Linguistics and Philosophy, 8:333–343.

... for more, check out: https://www.aclweb.org/anthology/W/W14/W14-5310.pdf tankä :)