Language technology for low-resource languages

Day 4/5

LOT 2018, Groningen

Yves Scherrer, University of Helsinki

Data transfer and model transfer

- How do the different methods handle the following characteristics?
  - Word order differences between HRL and LRL
  - Polysemy/homonymy in HRL/LRL
  - False friends
  - Grammatical categories not present in HRL/LRL

Closely related languages

- If the LRL is closely related to the HRL, how does this simplify...
  - ... the use of data transfer methods?
  - ... the use of model transfer methods?

Closely related languages

- Data transfer is easier
  - Parallel corpora for projection are “more parallel”
  - Training data translation can be done word-by-word or character-by-character
- Model transfer works better
  - Larger number of identical words, so plain model transfer works
  - Assumptions about constant word order are more realistic
  - Cognate matching instead of bilingual lexicon induction

Case studies

- Training data translation
  - Czech -> Slovak
- Test data backtranslation
  - Modern Slovene <- historical Slovene
  - Standard German <- Alsatian
  - Standard German <- Swiss German
- Plain model transfer
  - Slovak, Polish, Ukrainian, Russian -> Rusyn
- Relexicalization
  - Cognate matching, various language configurations
- Delexicalization with character embeddings
  - Various language configurations

Treebank translation for closely related languages

- VarDial 2017 shared task on cross-lingual parsing
  - Take a Czech treebank
  - Translate it to Slovak word by word
    - No unaligned words, no dummy nodes to process
    - Borders between treebank translation and relexicalization are somewhat blurry…
  - Train a parser on the “slovakified” data

	LAS	UAS
Plain model transfer	53.72	65.70
Treebank translation	64.05	73.16
Supervised model	69.14	76.57

Treebank translation for closely related languages

- Universal dependency annotations are not harmonized (enough) across languages
  - The Czech treebank has a *nummod:gov* dependency relation, the Slovak dev data only has *nummod*.
  - The Czech treebank distinguishes *DET* and *PRON*, the Slovak dev data does not.
- Also train a new Slovak POS tagger on translated data to provide the parser with consistent input features

Treebank translation for closely related languages

- Results:

	LAS	UAS
Plain model transfer	53.72	65.70
Treebank translation	64.05	73.16
Treebank translation with harmonization and cross-tagging (+ other improvements)	78.12	84.92
Supervised model	69.14	76.57

Treebank translation for closely related languages

- Using a cross-lingual transfer method yields better results than a supervised model!
  - Czech treebank: 68 000 sentences
  - Slovak treebank: < 10 000 sentences

- Similar, although not as spectacular results for Croatian and Norwegian

M. Zampieri et al. (2017): *Findings of the VarDial Evaluation Campaign 2017.* Proceedings of VarDial 2017.
Rudolf Rosa et al. (2017): *Slavic Forest, Norwegian Wood*. Proceedings of VarDial 2017.

Test data backtranslation for historical varieties

- Historical Slovene text from 1750-1900
- How can we tag/lemmatize the historical data?
  - Tagger and lemmatizer for modern Slovene are available
  - Lemmatization = associate historical word form with modern lemmas

Y. Scherrer Y & T. Erjavec (2016): Modernising historical Slovene words. *Natural Language Engineering* 22(6), 881-905.

Test data backtranslation for historical varieties

- Original:
  - Plain model transfer, apply modern models to historical data without modification
- Normalization:
  - Apply some rules to transform historical spelling to modern spelling, then apply modern models
- Modernization:
  - Learn character-level machine translation model on parallel data, then modernize all historical data, and apply modern models

Test data backtranslation for historical varieties

- Three sub-periods:
  - 1750-1800
  - 1800-1850 (pre spelling reform)
  - 1850-1900 (post spelling reform)

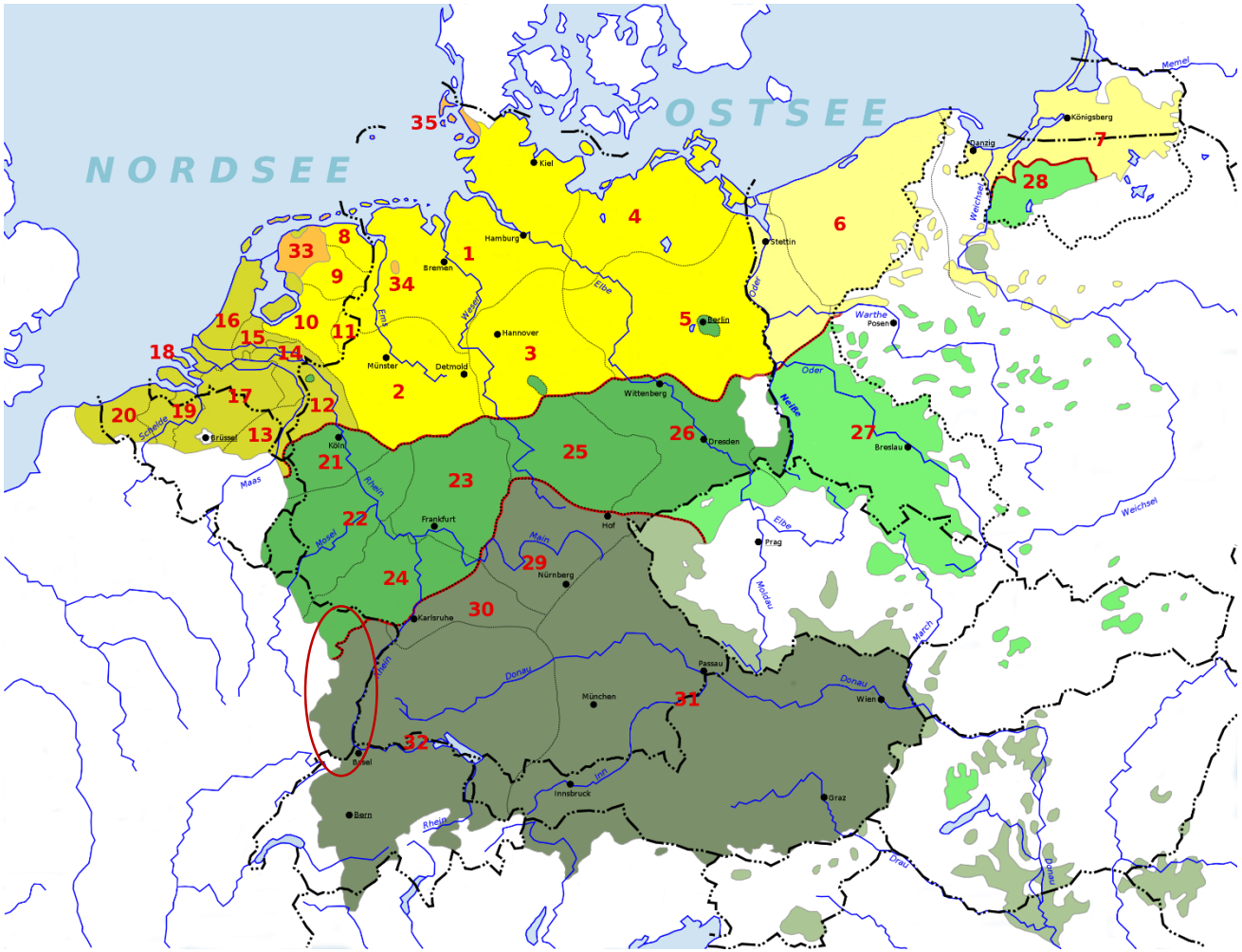
1790	Al	ta	nar bòl	vashna	reſsniza	je	moja	lubęsen	prut	Nęshki.
(18B)	*ali*	*ta*	*najbolj*	*važna*	*resnica*	*je*	*moja*	*ljubezen*	*proti*	*nežki*
1843	poboshnim	ſerzam	in	veſtjo	pridnoſt	in	ljubesin	k	ſvojimu	ſtanu	sdrushi
(19A)	*pobožnim*	*srcem*	*in*	*vestjo*	*pridnost*	*in*	*ljubezen*	*k*	*svojemu*	*stanu*	*združi*
1872	Otroška	ljubezen	naj	zmír	te	navdaja	Za	starše,	za	brate,	Bogá	in	cesarja
(19B)	*otroška*	*ljubezen*	*naj*	*zmeraj*	*te*	*navdaja*	*za*	*starše,*	*za*	*brate,*	*boga*	*in*	*cesarja*

Test data backtranslation for historical varieties

- Three sub-periods:
  - 1750-1800
  - 1800-1850 (pre spelling reform)
  - 1850-1900 (post spelling reform)

	Tagging			Lemmatization		
	1750	1800	1850	1750	1800	1850
Original	27.9	53.0	80.8	7.7	17.2	80.4
Normalized	41.6	72.3	81.5	22.7	57.3	82.1
Modernized	64.1	77.0	82.3	57.4	68.5	84.3

Test data backtranslation Alsatian



Alsatian

- German dialect (High German / Alemannic) spoken in France
- Several dialectal variants, no standardized spelling
- Minority language (normative pressure from French), few available resources

Delphine Bernhard & Anne-Laure Ligozat: Hassle-free POS-Tagging for the Alsatian Dialects. In Marcos Zampieri & Sascha Diwersy: *Non-Standard Data Sources in Corpus Based-Research*, Shaker, 85-92, 2013, ZSM Studien.

Alsatian

Ìch hà nìtt g'nüä Gald.
Ìch hàbb nìtt gnüej Gald.
Ìch hàbb nìtt genüej Gald.
Ìch hàbb nìtt gnüej Gël̈d.
Ìch hàbb nìtt genüej Gël̈d.

Ich habe nicht genug Geld.
I don't have enough money.

German	English	Alsatian variants
Küche	kitchen	Kuch, Kucha, Kische, Khésche, Kùch, Kücha, Kuche, Kiche, Kuchi
Montag	monday	Mondàà, Mantig, Mandig, Mondàà, Mondoe, Mondàj, Maandi, Mandi

Tagging Alsatian

- Plain model transfer:
  - Apply Standard German tagger on Alsatian data
- Example:
  - Words in red are identical in Standard German

Brüchsch	kenn	Angscht ze han	for	mich , papa .

Text	TreeTagger	Stanford Tagger
Alsace [news]	48%	56%
Duttlenheim [theater]	67%	77%
Hoflieferant [theater]	64%	67%
Wikipedia	50%	53%

Tagging Alsatian

- Translate the 107 most frequent Alsatian word forms to German:

Brüchsch	kenn	Angscht	ze han	for	mich , papa .
Brüchsch	keine	Angscht	zu haben	für	mich , papa .

Text	TreeTagger	Stanford Tagger	Translated tokens
Alsace	79% (+31%)	86% (+30%)	32%
Duttlenheim	86% (+19%)	88% (+11%)	19%
Hoflieferant	78% (+14%)	82% (+15%)	17%
Wikipedia	83% (+33%)	85% (+32%)	36%

Tagging Alsatian

- Plain model transfer is not enough
  - Even when languages are closely related
  - 48%-77% accuracy for tagging
- Manual annotation (translation) of about 100 high-frequency words helps a lot:
  - 48% → 79%, 77% → 88% accuracy
  - Recall unrelated languages:
    2h of annotation ≈ 1000 word/tag pairs
  - The most frequent words help most. Influence of additional words decreases with decreasing frequency.
  - Not only the translated word forms are tagged more correctly, also their surroundings.

Tagging Swiss German

- The ArchiMob corpus:
  - Transcribed video interviews in various Swiss German dialects
  - Goal: part-of-speech tagging
- Available training resources:
  - TüBa-D/S: spontaneous dialogues, Standard German, 360k tokens
  - NOAH: Swiss German news articles, not the same transcription guidelines, 73k tokens
- What is the best way to tag the ArchiMob data?

T. Samardzic et al. (2016). *ArchiMob - A Corpus of Spoken Swiss German.* Proceedings of LREC 2016.

Tagging Swiss German

- Normalization: learn a character-level machine translation on hand-normalized data

BTagger trained on TüBa-D/S and NOAH
Applied to original or normalised forms

Train	Test	% Acc.	% OOV
TüBa-D/S	Normalised	70.31	24.21
NOAH	Original	60.56	30.72

Tagging Swiss German

- We don't need to learn how to tag punctuation signs, as they don't occur in the ArchiMob transcriptions…

BTagger trained on TüBa-D/S and NOAH
Applied to original or normalised forms

Train	Test	% Acc.	% OOV
TüBa-D/S	Normalised	70.31	24.21
NOAH	Original	60.56	30.72
Removed punctuation:			
TüBa-D/S	Normalised	70.68	24.21
NOAH	Original	73.09	30.72

Tagging Swiss German

- This is a negative result for cross-lingual learning methods…
  - Using test data backtranslation (=normalization) does not work better than a supervised model, even though the genre and transcription guidelines of the supervised model are different!

Tagging Swiss German

- Further improvements:
  - Annotate a document
  - Hand-correct the annotations
  - Add the corrected document to the training data
  - Train a new tagger and repeat

Round	1	2	3	4
Accuracy	79.99	84.08	88.55	90.09

Normalizing Swiss German

- Normalization may not be useful for tagging, but can we use it for other purposes?
  - Dialect-independent search

Query üüs **169** (1,424.78 per million)

Page 1 of 9 Go Next | Last

#10	<file> <s> <align> jò frau walser / chönd si	üüs	/uns	sège wo si ufgwachse sind </align> </s>
#1029	händ wellen iizie bis zu dèm phunggt mönds	üüs	/uns	der erloo / das zalemer nüüd / und das
#1911	</s> <s> <align> näi das isch nattüürli für	üüs	/uns	e käis gsii mit puurne häts e käi aarbäitsloosikäit
#3464	nüd / eh / ja kwaasi gchöörsch nüd zu	üüs	/uns	/ neh </align> </s> <s> <align> mh / händ
#4196	daas hät de vilicht daas uusgmacht das men	üüs	/uns	dän ebe d herepuuren aaghänggt hät / sii
#4470	äifacht nüd gläge gsìi ùnd das (isch an)	üüs	/uns	überggangen äigentlich / ich ha düch mängmal
#4775	nocher) emal daa hindere zûglet und die hät	üüs	/uns	dän e foorm praacht / wù mer uf em gaasröschscho
#5233	</align> </s> <s> <align> ja / me hät bi	üüs	/uns	isch am tischsch polittisiert woorde /
#5290	dem turner wù dän ùnd dem schräiner wùn	üüs	/uns	de deet i dere ziit chù isch chù hälffe
#5950	soo jung chinderleemig gchaa und der isch	üüs	/uns	daa chùù gù mälche / ùnd mir händ ja das
#9115	truppe nööd / (da) isch natüürli füür	üüs	/uns	es eräignis gsii as daa plözlich eson e
#9633	dän aber beedi chänne säge si chänd ez mit	üüs	/uns	chùù mir wüssed wo die sind mer händ der
#14832	di zwäihundert gsìi wümer gchaa händ won	üs	/uns	en aart ver / verchäufft hetted / und zwaar
#18861	hindere güù / gù fare / und / eh / da simmer	üüs	/uns	daa maal begägnet und händ dän daa halt
#20561	si sägid hüt na / mir gend hai wen si zu	üüs	/uns	chemid / jaa ai schwöschter isch uisgwanderet
#20888	und der isch öü zur chilen uis und hinder	ïs	/uns	nachen und wommer is huis ine sind ischsch
#20898	nachen und wommer is huis ine sind ischsch er	ïs	/uns	nachecho / glüüte / und gsait gäl du bisch
#20981	scho vili jaar ghüraate gsii / (het) mer	ïïs	/uns	gschpaart und he / hend emal welle die
#20998	gschwüschterti psueche / di ainte die hend	ïïs	/uns	gschribe mir zaalidich d rais wenner nur
#21875	/ mir hend der vatter acht jaar de na bi	ïïs	/uns	ghaa und im maa si mueter (12:10) / maa

Page 1 of 9 Go Next | Last

Normalizing Swiss German

- Normalization may not be useful for tagging, but can we use it for other purposes?
  - We can create a distinct normalization system for each document
  - The normalization systems are character-level statistical machine translation systems, so transformations are stored in phrase tables
  - Hypothesis: certain phrase pairs show regionally different frequency distributions

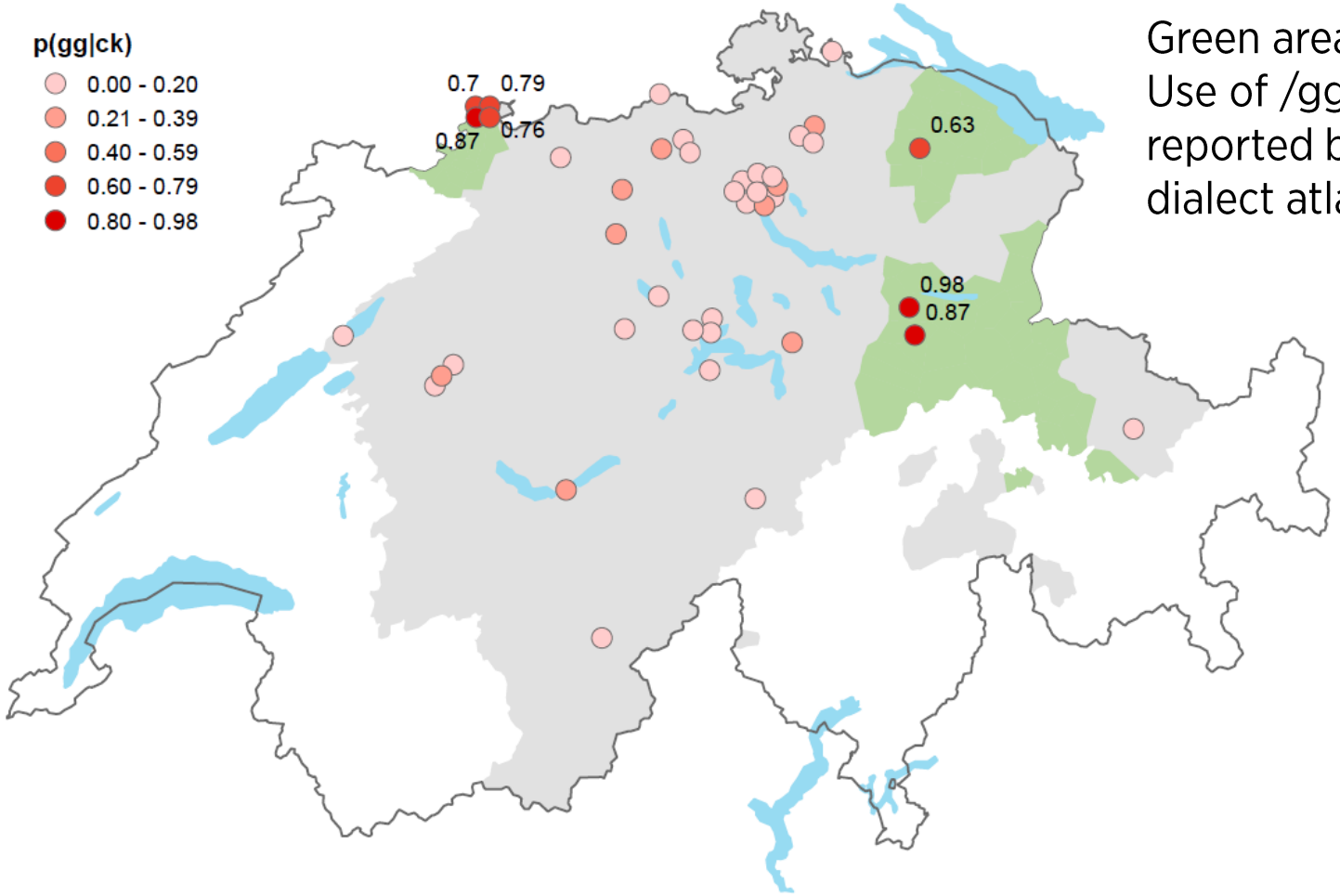
Normalizing Swiss German

- From which dialectal transcriptions arises normalized *ck*?
- Look for $p(* \,|\mathrm{ck})$ in phrase tables:

Transcription -> Normalization	Document 1	Document 2
k -> ck	37.0%	95.2%
gg -> ck	63.0%	2.4%
ch -> ck	0%	2.4%

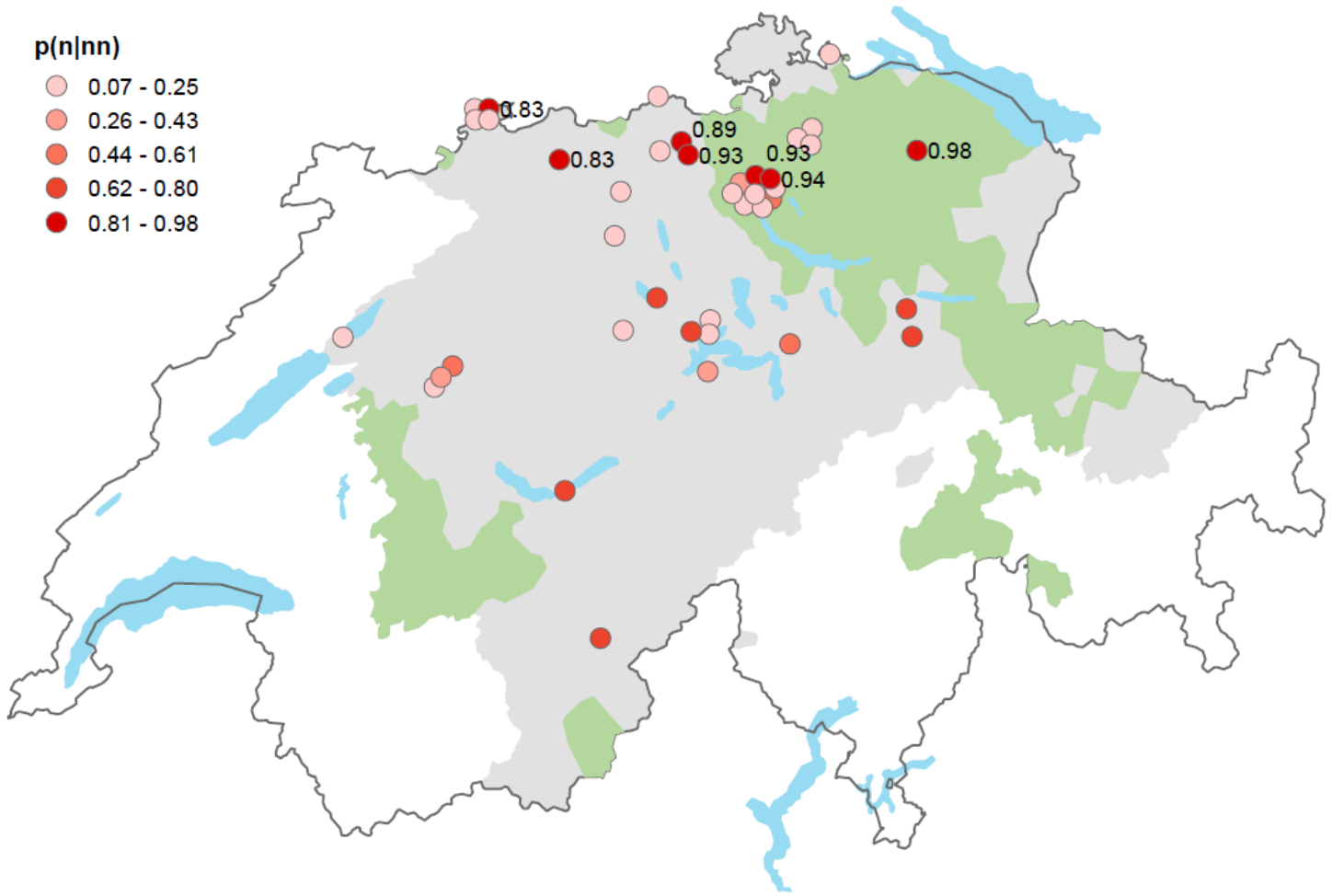
- Plot probabilities for one variant, e.g. $p(\mathrm{gg}|\mathrm{ck})$

Normalizing Swiss German

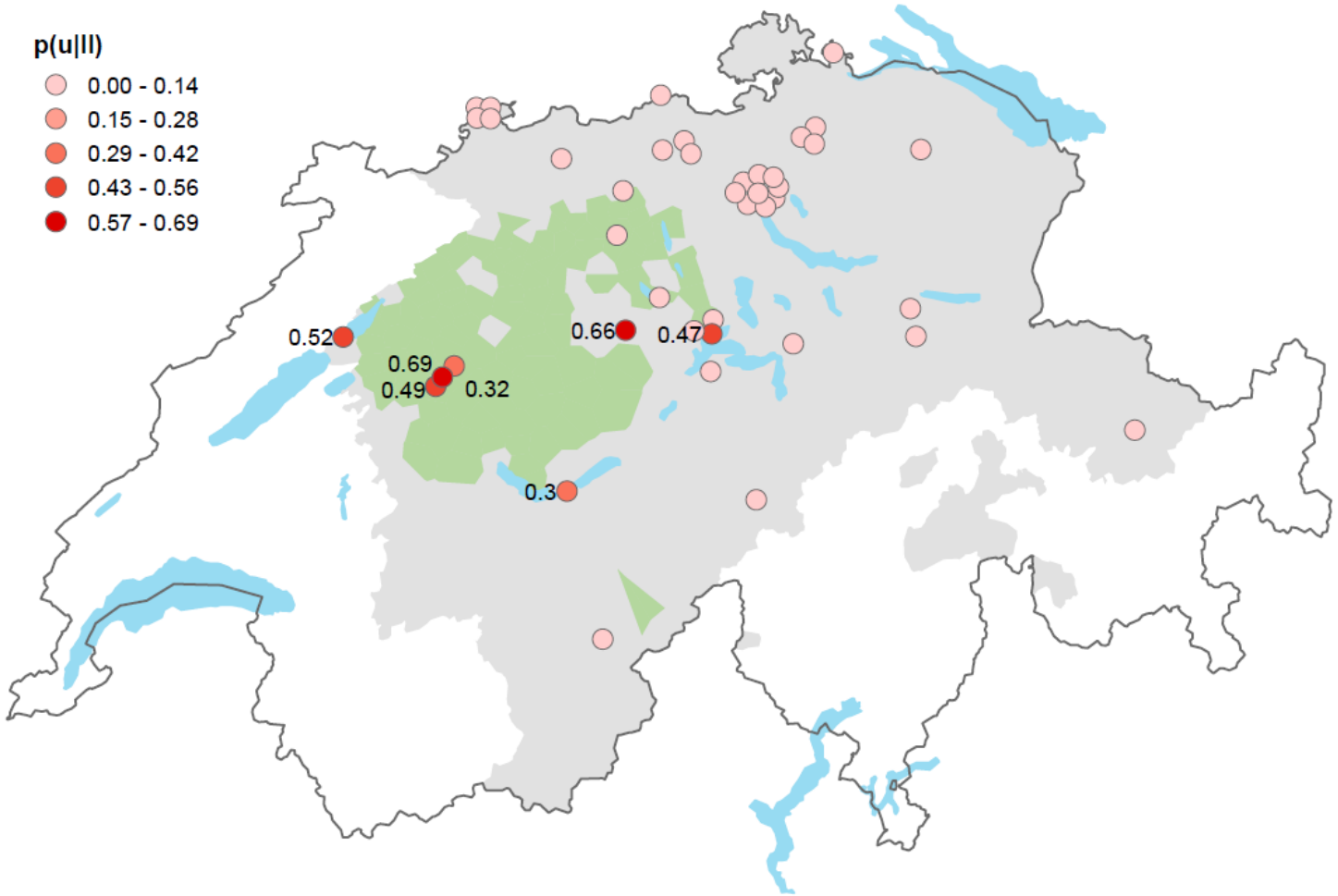


Green areas:
Use of /gg/
reported by a
dialect atlas

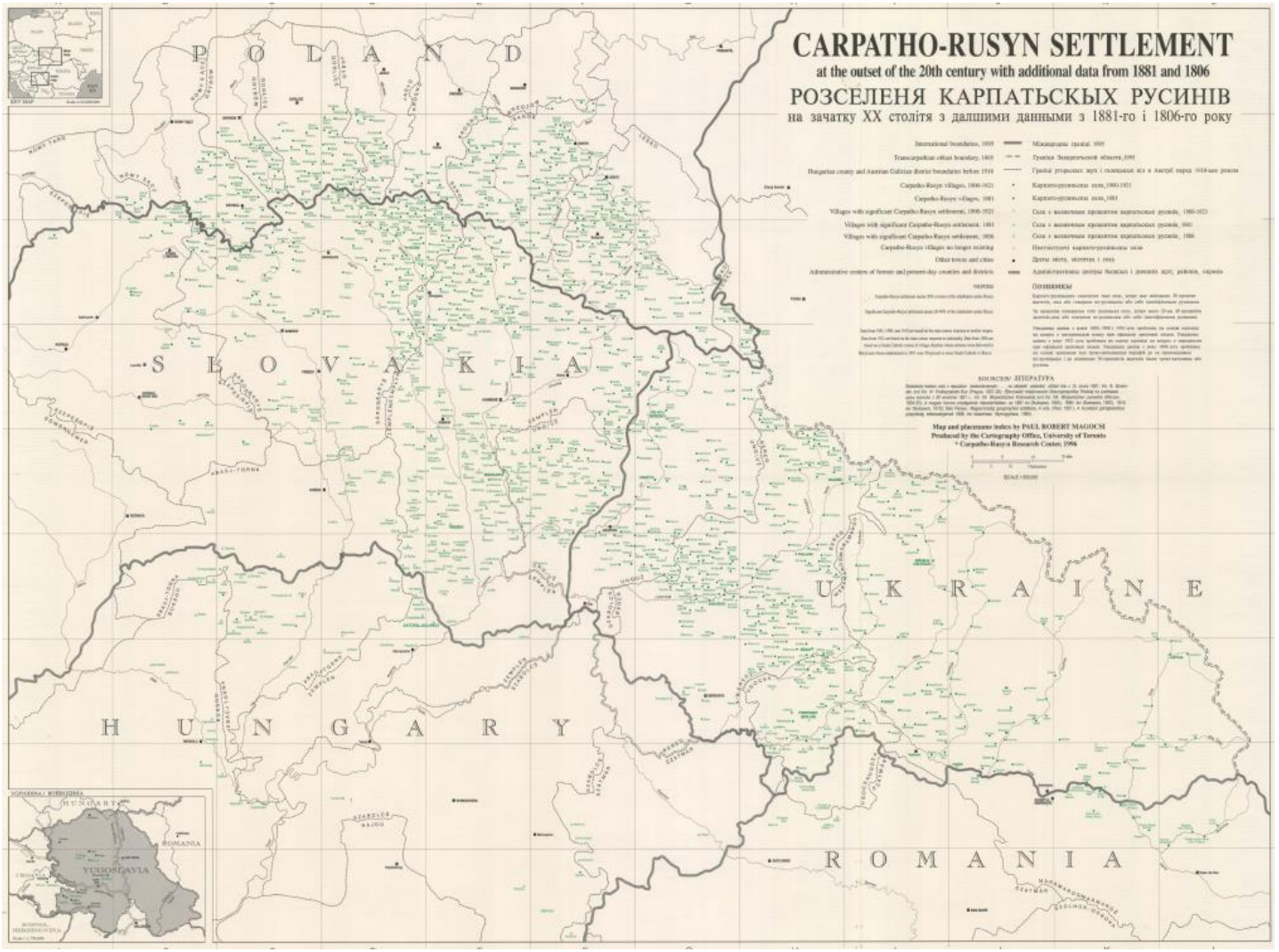
Normalizing Swiss German



Normalizing Swiss German



Multi-source tagging for Rusyn



Multi-source tagging for Rusyn

- Rusyn: Slavic minority language, several regional dialects
- Corpus of Spoken Rusyn:
  - Transcribed speech from different areas
  - Written in Cyrillic script
  - To be annotated: morphosyntactic tagging

но	безівно	як	він	дезь	го	убили	у	вовках	.
INTJ	ADV	PRON	PRON	PRON	PRON	VERB	ADP	NOUN	PUNCT
		PronType=Int	Case=Nom	PronType=Int	Case=Acc	Mood=Ind		Case=Loc	
			Gender=Masc		Gender=Masc	Number=Plur		Gender=Masc	
			Number=Sing		Number=Sing	Tense=Past		Number=Plur	
			Person=3		Person=3				
			PronType=Prs		PronType=Prs				

Yves Scherrer & Achim Rabus (2017). *Multi-source morphosyntactic tagging for Spoken Rusyn*. In Proceedings of VarDial 2017.

Data

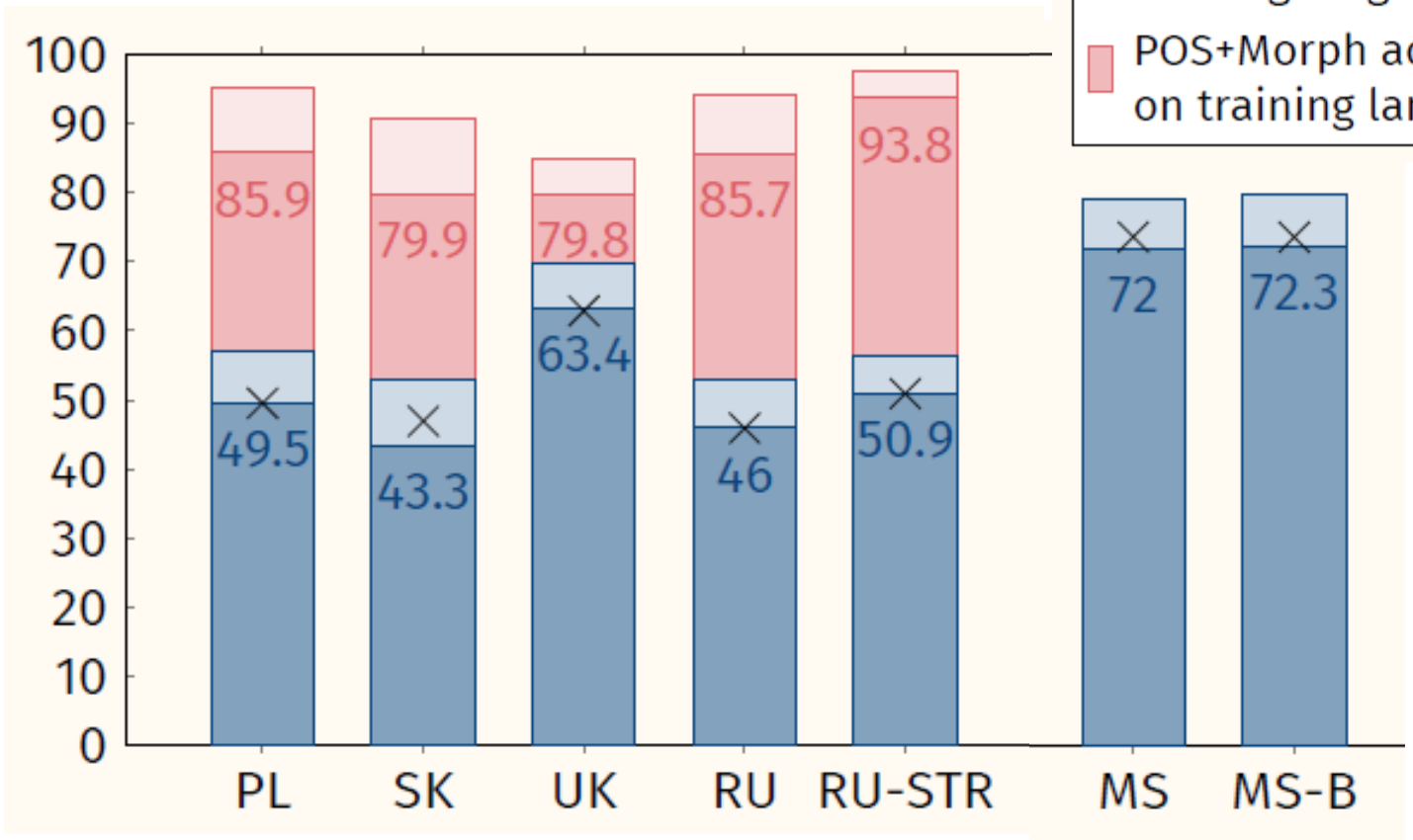
Main idea: train tagger on data from etymologically related neighboring languages: Ukrainian, Polish, Slovak, Russian.

Language	Origin	Training data			Test data		
		Sent.	Tokens	Tags	Sent.	Tokens	Tags
PL	UD 1.4	6 800	69 499	920	700	6 887	448
RU	UD 1.4	4 029	79 772	704	502	10 044	410
RU-STR	UD 1.4 SynTagRus	48 171	850 689	580	6 250	109 694	501
SK	UD 1.4	8 483	80 575	657	1 060	12 440	426
UK	UD 1.4 + additional	4 162	71 580	1 040	55	395	92
RUE1	Manually annotated Rusyn gold standard				104	1 050	96
RUE2	Unannotated Corpus of Spoken Rusyn				5 922	75 201	—

- Polish and Slovak data are cyrillicized using a set of rewrite rules.
- We use the MarMoT tagger (Müller et al. 2013) with the morphological tagging option for all experiments.

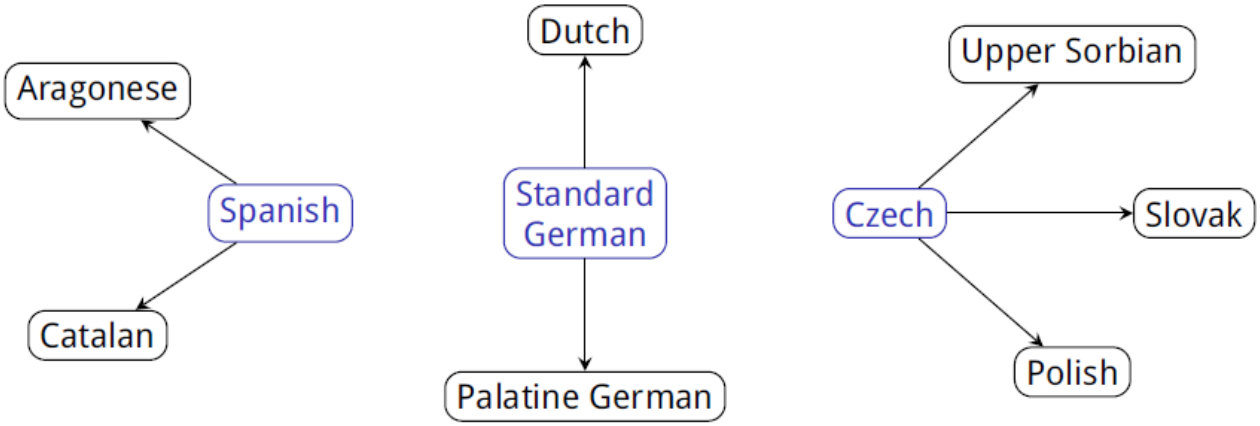
Results

MS: Concatenate training data from PL, SK, UK, RU, RU-STR, train single tagger
MS-B: with added word clusters



Relexicalization with cognate matching

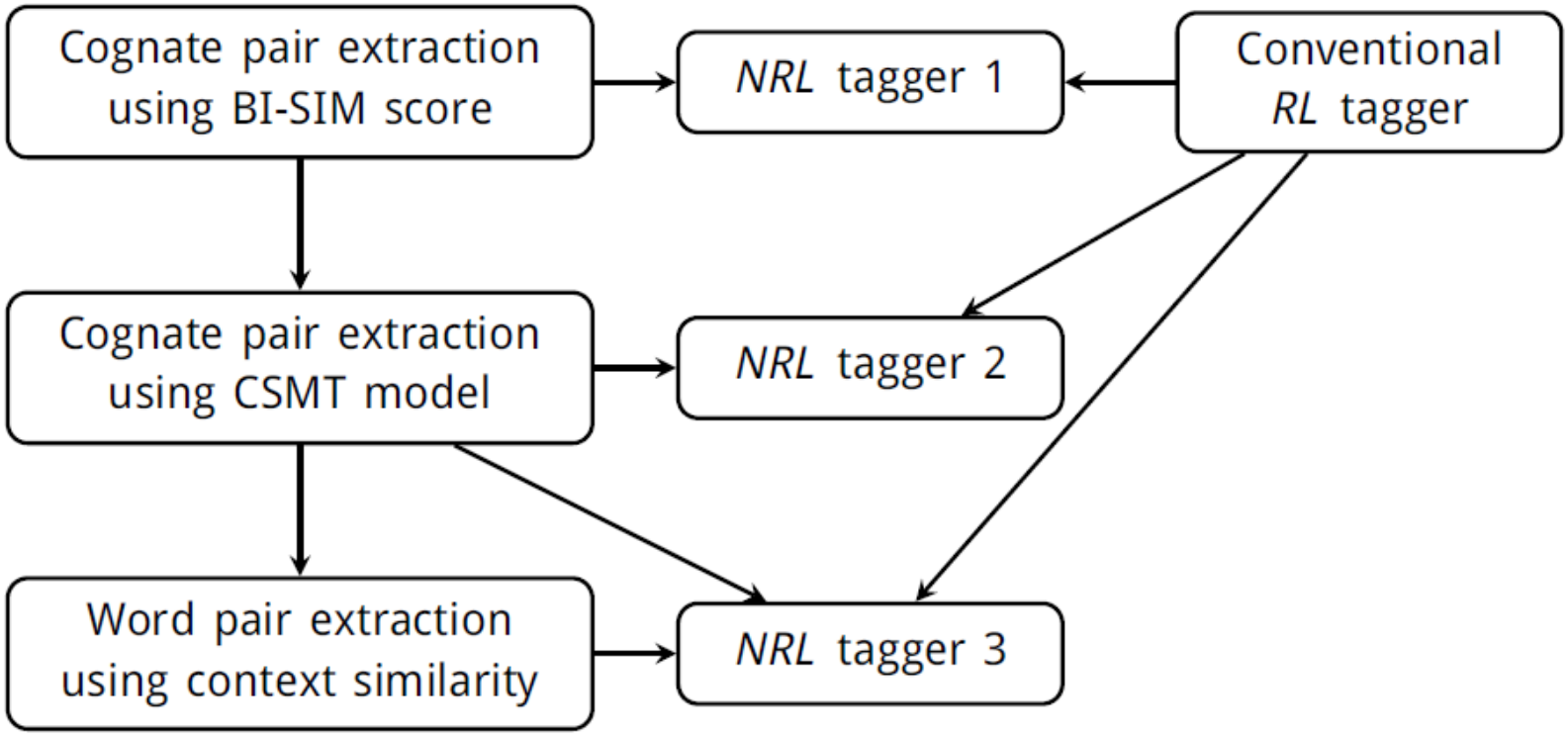
- Goals:
  - Don’t use any bilingual resources (dictionaries, parallel corpora)
  - Don’t do any manual annotation work
  - But restrict experiments to closely related languages
  - Comparable raw text for all languages is available (Wikipedia)



Cognate matching

- Approach 1 (BI-SIM):
  - For each word of the HRL, find the most similar LRL word, then relexicalize the HRL tagger
- Approach 2 (CSMT):
  - Train a model of language relatedness with the word pairs from approach 1
  - For each HRL word, find the most similar LRL word according to this model, and relexicalize the HRL tagger
- Approach 3 (Context):
  - Train contextual similarity model using the word pairs from approach 2 as seed lexicon
  - Find additional word pairs with contextual similarity and relexicalize the HRL tagger

Cognate matching

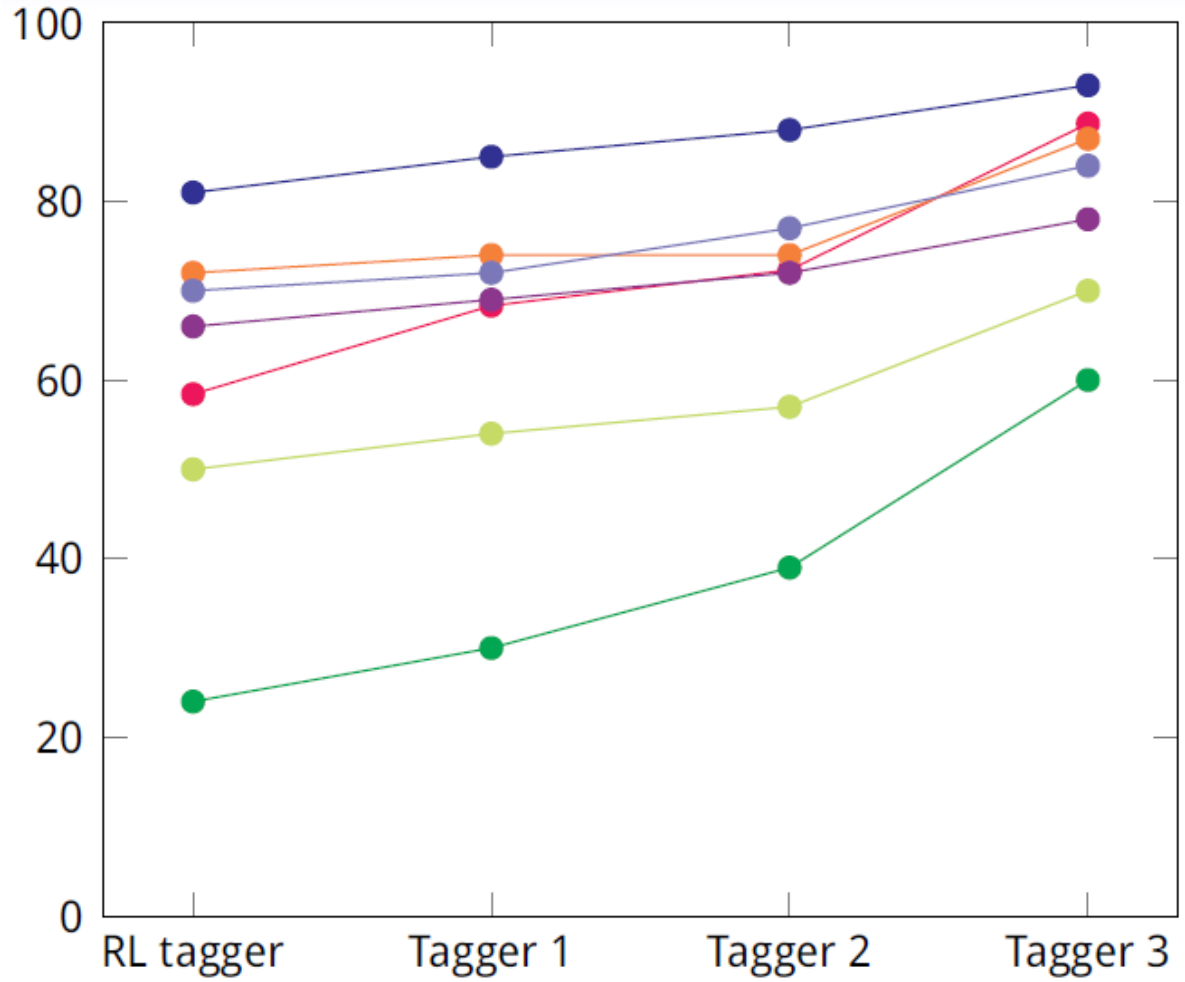


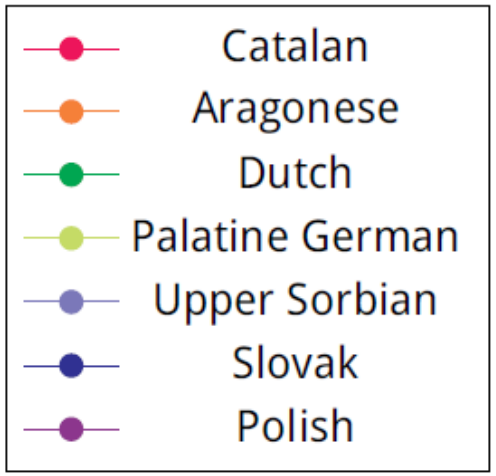
Y. Scherrer & B. Sagot (2014). *A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages*. In Proceedings of LREC.

Examples (Dutch-German)

Dutch	German candidates			Correct German
	retained by BI-SIM	CSMT results	retained after filtering	
vegetatie	vegetative vegetation	vegetation	vegetation	vegetation
groenen	grossen großen groben	groenen grünen	grünen	grünen
amfibieën	—	amphibien	amphibien	amphibien
enkel	enkel onkel	enkel	enkel	einfach (false friends)
zweden	zwecken zweigen	zweiten schweden	zweiten	schweden (wrong)

Cognate matching – Results





Test data

Catalan: AnCora corpus (500 000 words)

Other languages: manually annotated Wikipedia excerpts (30-100 sentences per language)

Cognate matching

- Monolingual datasets

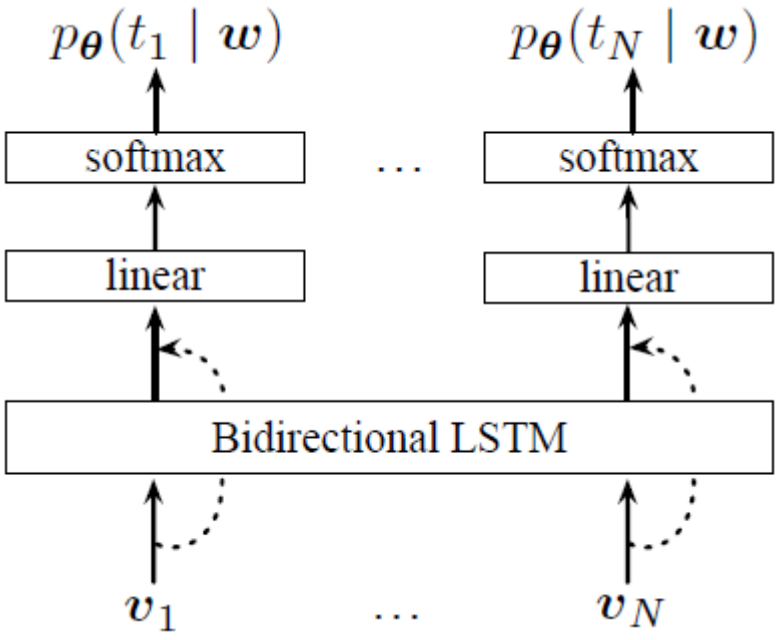
LRL	Tokens	HRL	Tokens
CA	140M	ES	432M
AN	5.4M		
NL	0.5M	DE	613M
PFL	0.3M		
HSB	0.9M	CS	86M
SK	30M		
PL	206M		

Cognate matching

- Cognate matching needs to be
  - language-pair-dependent
    (approach 2 vs approach 1)
  - enhanced by semantic information
    (approach 3 vs approach 2)
- Problems:
  - Ambiguity handling:
    assumed 1-1 correspondence for relexicalization
  - Tokenization differences:
    assumed 1-1 correspondence for relexicalization
  - Germanic languages perform poorly:
    compounds? spelling? data size?

Neural network tagging and parsing

- Neural network models automatically perform some kind of delexicalization
- First layer:
  - Lookup table
  - Word form -> embedding vector
- The vs are language-independent
- Delexicalization:
  - Replace lookup table by bilingual embeddings

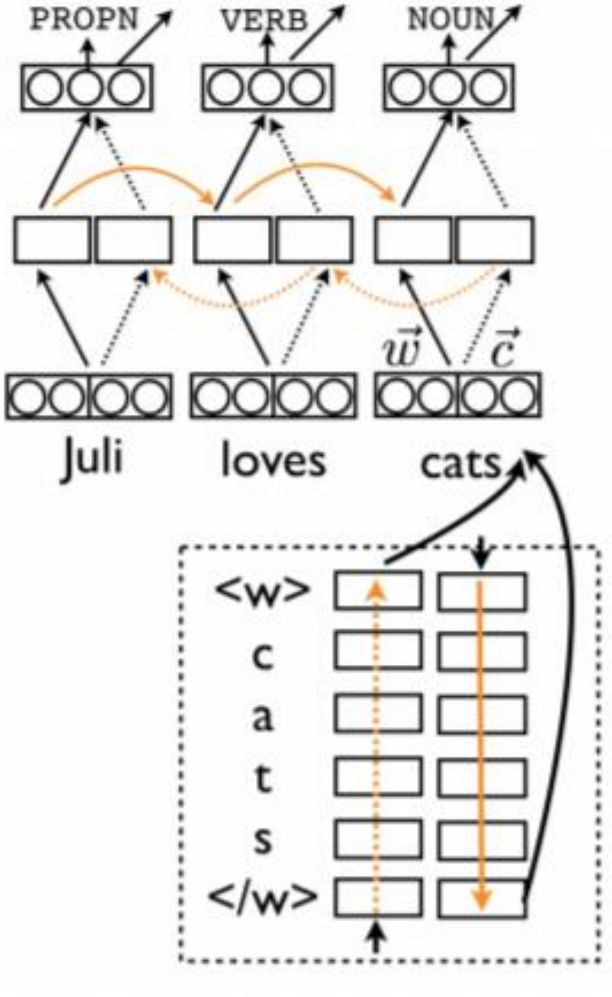


What type of word embeddings?

- Standard word embeddings
  - Each distinct word form is considered an atomic entity that gets its own embedding vector based on context
- Character-level word embeddings / "character embeddings"
  - Each word form is split into its characters
  - A recurrent or convolutional network is trained to predict the character sequence
  - A summary of the character embeddings (e.g. pooling) is fed to the subsequent layers
  - Such embeddings represent word-internal structure

What type of word embeddings?

- Common wisdom:
  - Both types of embeddings are helpful because they encode different aspects of the data
  - Concatenate both vectors
- What about cross-lingual transfer models?
  - Cross-lingual word embeddings require large corpora (+ dictionaries)
  - What about character embeddings?



Impact of different embedding types

On average, character embeddings are much more helpful than word embeddings

Language	w	c	w + c	Language	w	c	w + c
ar	95.48	98.68	98.91	he	93.97	93.74	95.79
bg	95.12	97.89	98.02	hi	95.99	93.4	96.23
cs	93.77	96.38	97.8	hr	89.24	95.32	94.76
da	91.96	95.12	96.19	id	90.48	91.37	93.11
de	90.33	90.02	92.64	it	96.57	95.62	97.59
en	92.1	91.62	94.46	nl	84.96	89.11	93.32
es	93.6	93.06	95.12	no	94.39	95.87	97.57
eu	88	92.48	94.7	pl	89.73	95.8	96.41
fa	95.31	95.82	97.19	pt	94.24	95.96	97.53
fi	87.95	90.25	94.85	sl	91.09	96.87	97.55
fr	94.44	94.39	95.8	sv	93.32	95.57	96.36

B. Plank, A. Søgaard, Y. Goldberg: *Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss.* ACL 2016.

Cross-lingual tagging with character embeddings

- Use related languages
- Adapt the HRL tagger to LRL by including a small number of annotated LRL sentences

R. Cotterell & G. Heigold (2017): *Cross-lingual Character-Level Neural Morphological Tagging*. Proceedings of EMNLP.

Cross-lingual tagging with character embeddings

Romance				Slavic			
lang	train	dev	test	lang	train	dev	test
(ca)	13123	1709	1846	(bg)	8907	1115	1116
(es)	14187	1552	274	(cs)	61677	9270	10148
(fr)	14554	1596	298	(pl)	6800	7000	727
(it)	12837	489	489	(ru)	4029	502	499
(pt)	8800	271	288	(sk)	8483	1060	1061
(ro)	7141	1191	1191	(uk)	200	30	25
Germanic				**Uralic**			
lang	train	dev	test	lang	train	dev	test
(da)	4868	322	322	(et)	14510	1793	1806
(no)	15696	2410	1939	(fi)	12217	716	648
(sv)	4303	504	1219	(hu)	1433	179	188

Table 2: Number of tokens in each of the train, development and test splits (organized by language family).

Cross-lingual tagging with character embeddings

- Results for Romance languages (100 LRL training sentences):

source language	(ca)	(es)	(fr)	(it)	(pt)	(ro)
(ca)	—	87.9%	84.2%	84.6%	81.1%	67.4%
(es)	88.9%	—	85.5%	85.6%	81.8%	69.5%
(fr)	88.3%	87.0%	—	83.6%	79.5%	**69.9%**
(it)	88.4%	87.8%	84.2%	—	80.6%	69.1%
(pt)	88.4%	88.9%	85.1%	84.7%	—	69.6%
(ro)	87.6%	87.2%	85.0%	84.4%	79.9%	—

Cross-lingual tagging with character embeddings

- Results for Slavic languages (100 LRL training sentences):

source language	(bg)	(cs)	(pl)	(ru)	(sk)	(uk)
(bg)	—	47.4%	44.7%	67.3%	39.7%	57.3%
(cs)	57.8%	—	56.5%	62.6%	62.6%	54.0%
(pl)	54.3%	54.0%	—	59.3%	57.8%	48.0%
(ru)	**68.8%**	48.6%	47.4%	—	46.5%	**60.7%**
(sk)	55.2%	57.4%	54.8%	61.2%	—	49.3%
(uk)	44.1%	36.0%	34.4%	43.2%	30.0%	—

Cross-lingual tagging with character embeddings

- Results for Northern Germanic languages:

		(da)	(no)	(sv)
source	(da)	—	77.6%	73.1%
	(no)	83.1%	—	75.7%
	(sv)	81.4%	76.5%	—

- Results for Uralic languages:

		(et)	(fi)	(hu)
source	(et)	—	**60.9%**	**60.4%**
	(fi)	**60.1%**	—	60.3%
	(hu)	47.1%	48.3%	—

Cross-lingual tagging with character embeddings

- No parallel data, no dictionary mappings required
- However, small amount of LRL training data required
- Results not impressive
- Some missing baselines:
  - Train on only 100 LRL sentences, without HRL data
  - Train only on HRL data without LRL sentences
- Cross-lingual character embedding mappings?