Your noise is my research question! – Limitations of normalizing social media data

Torsten Zesch

Assistant Professor
Language Technology Lab, University of Duisburg-Essen

Associated Researcher
German Institute for International Educational Research, Frankfurt

Biographical Facts

Computer science background

2006 – 2012

Technische Universität Darmstadt (PhD/PostDoc)

- Semantic relatedness / Wikipedia

2012 – 2013

German Institute for International Educational Research (DIPF), Frankfurt

- Automatic scoring (PISA data)

Now

University of Duisburg-Essen, “Language Technology Lab”

Vice President of the German Society for Computational Linguistics & Language Technology (GSCL)

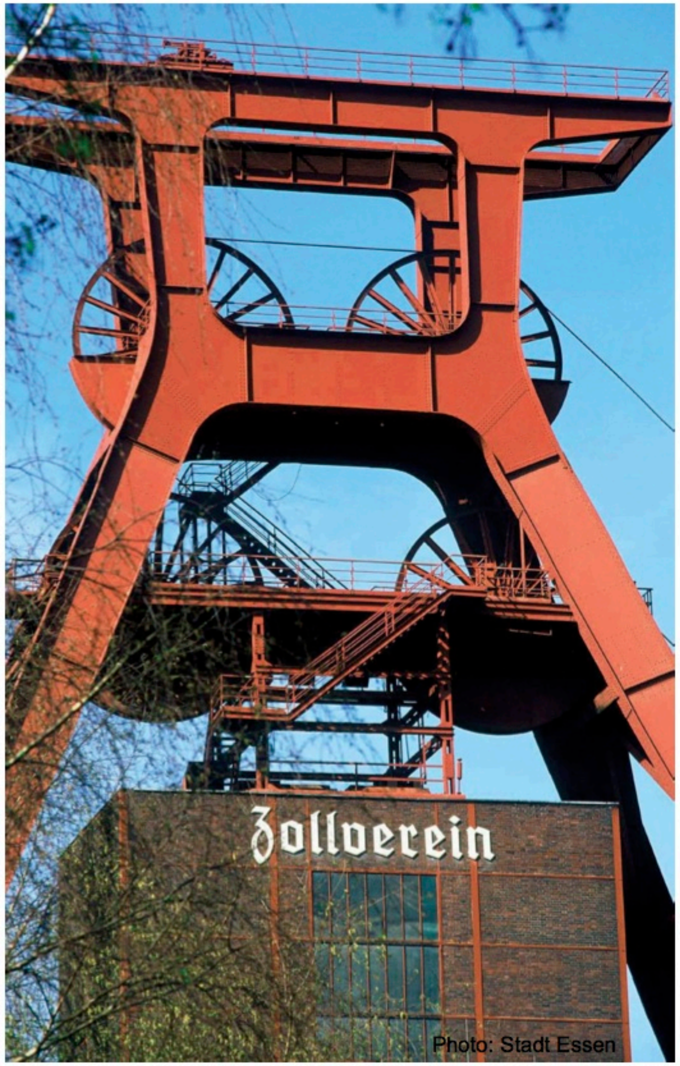
University of Duisburg-Essen

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

In the heart of Europe

University of Duisburg-Essen



Duisburg

- 490.000 inhabitants
- the most important steel production site in Europe
- logistical centre of Germany, with the largest inland port in Europe

Essen

- 570.000 inhabitants
- the cultural and economical centre of the Rhine-Ruhr region as well as a hotspot of the service industries

Research Interests

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

Language Technology for Education

- Automated scoring, exercise generation

Social Media Analysis

- Research training group "User centred Social Media"

Language Technology Infrastructure

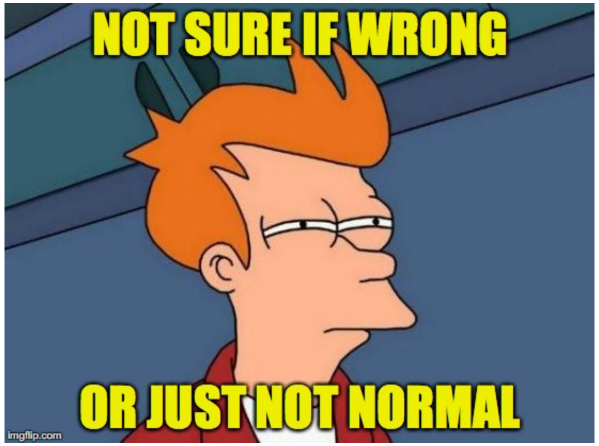
- Reproducibility & Replicability

Nrmliztion

Normalization Process

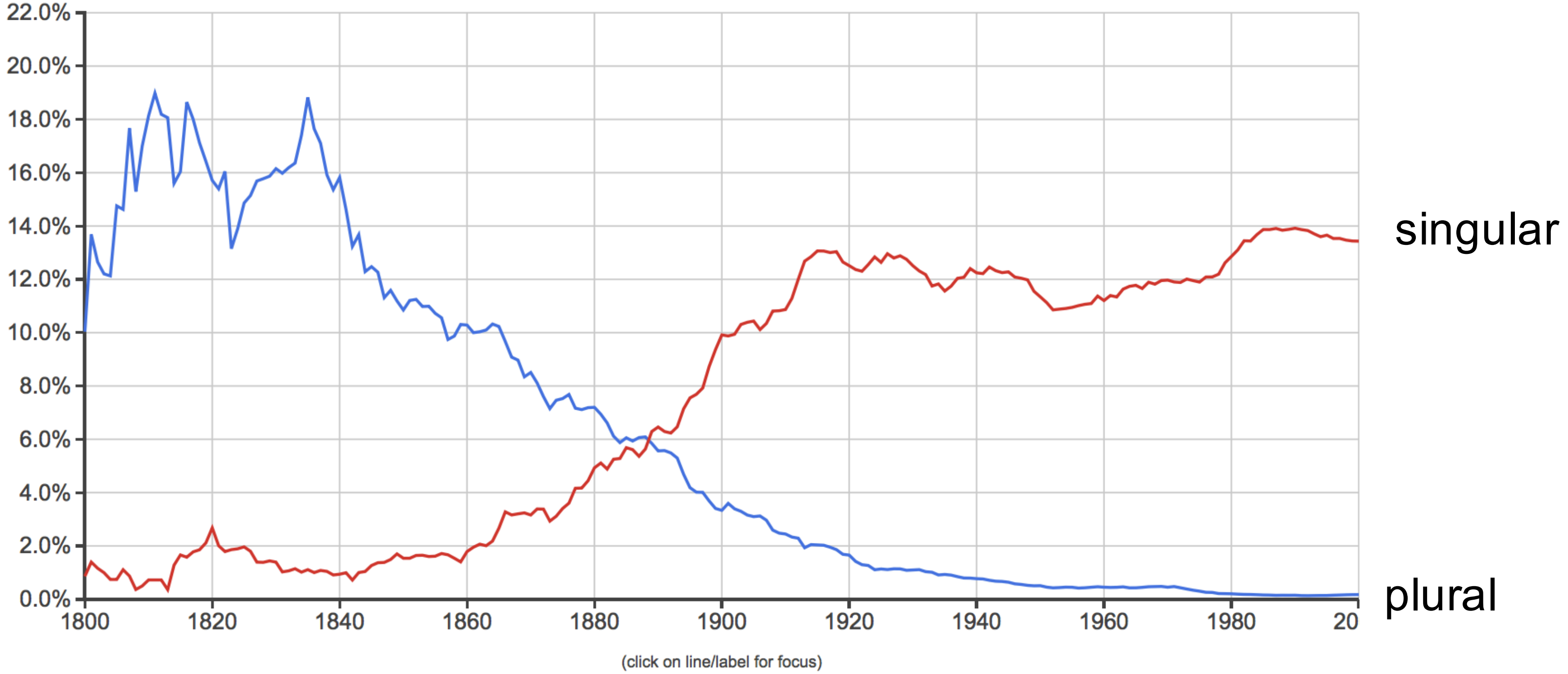
State A

State B

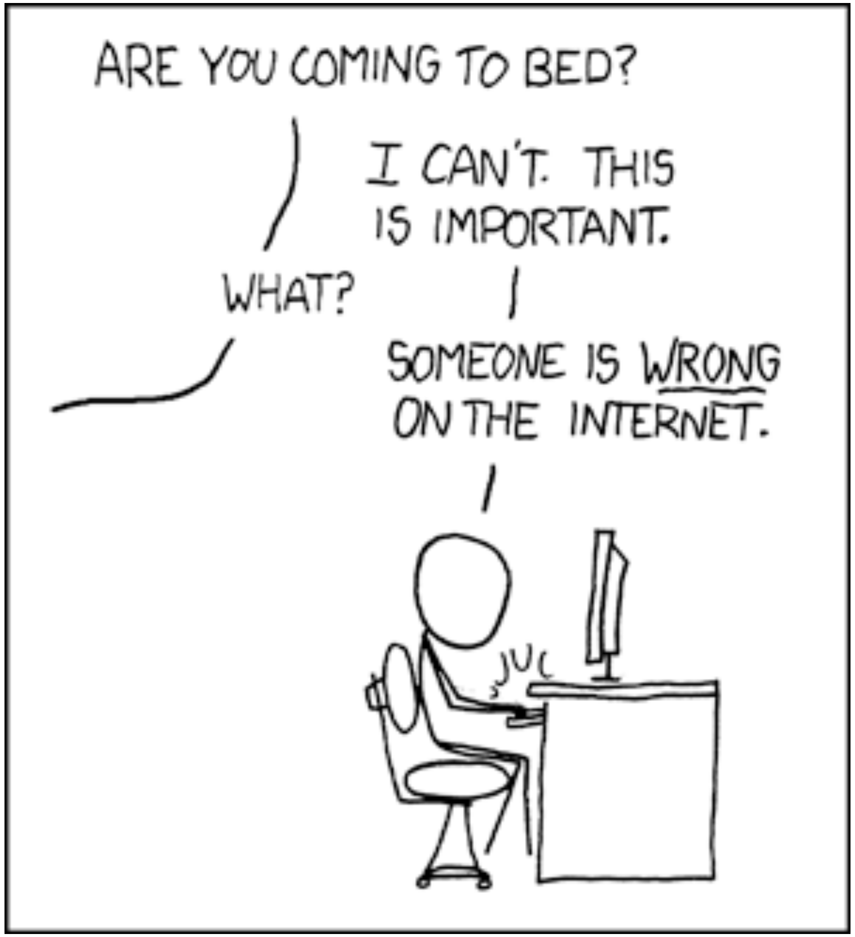


Dynamic Norms

United States (singular vs. plural)



Purpose of Normalization



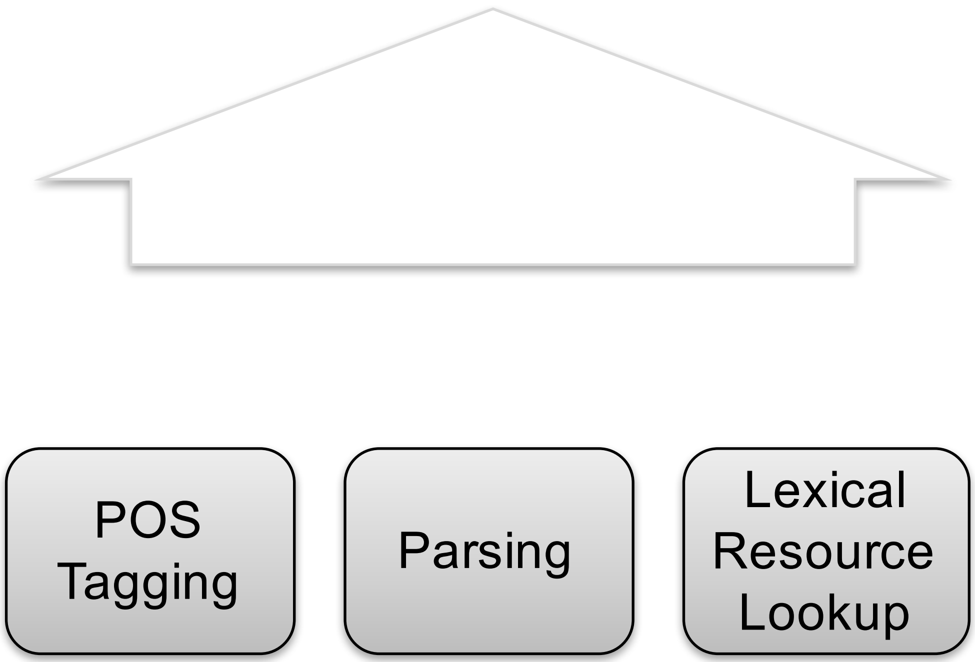
https://xkcd.com/386/

Use Cases for Normalization

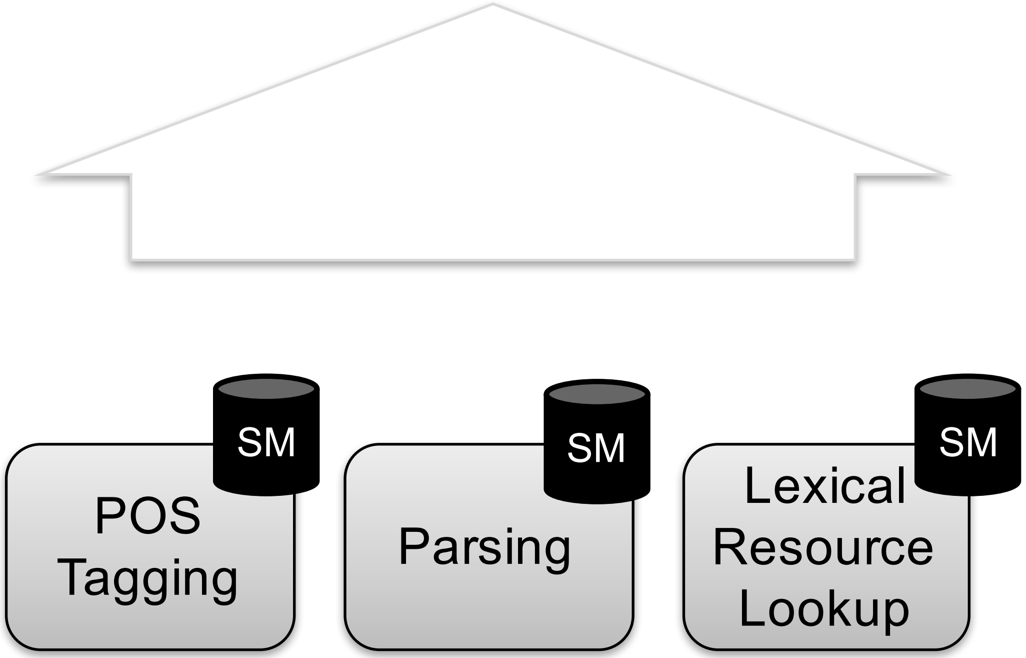
- Re-use existing tools / easier downstream processing
- Search / lookup
- Comparison / analysis

Standard Text

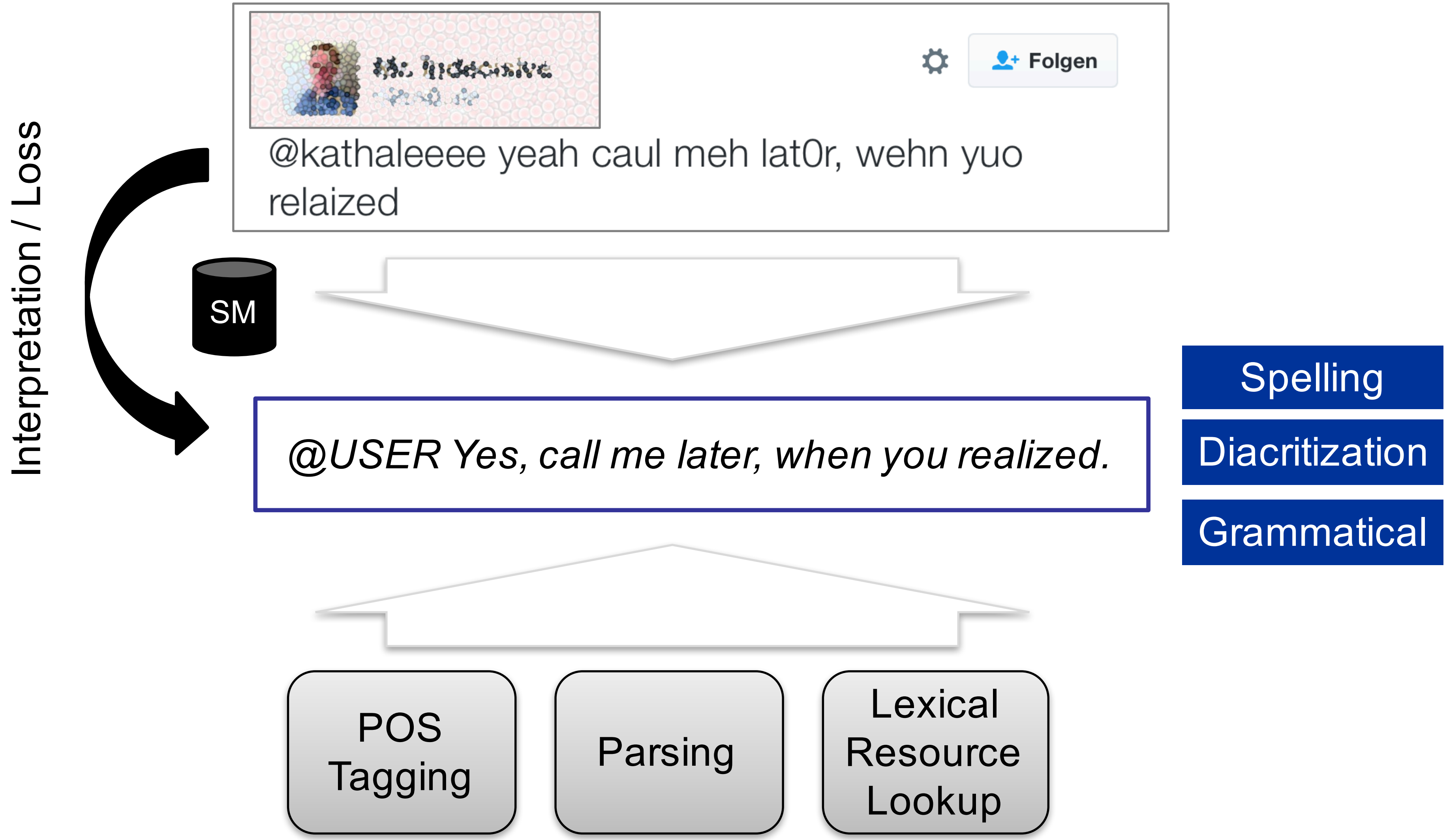
Please call me later when you have decided



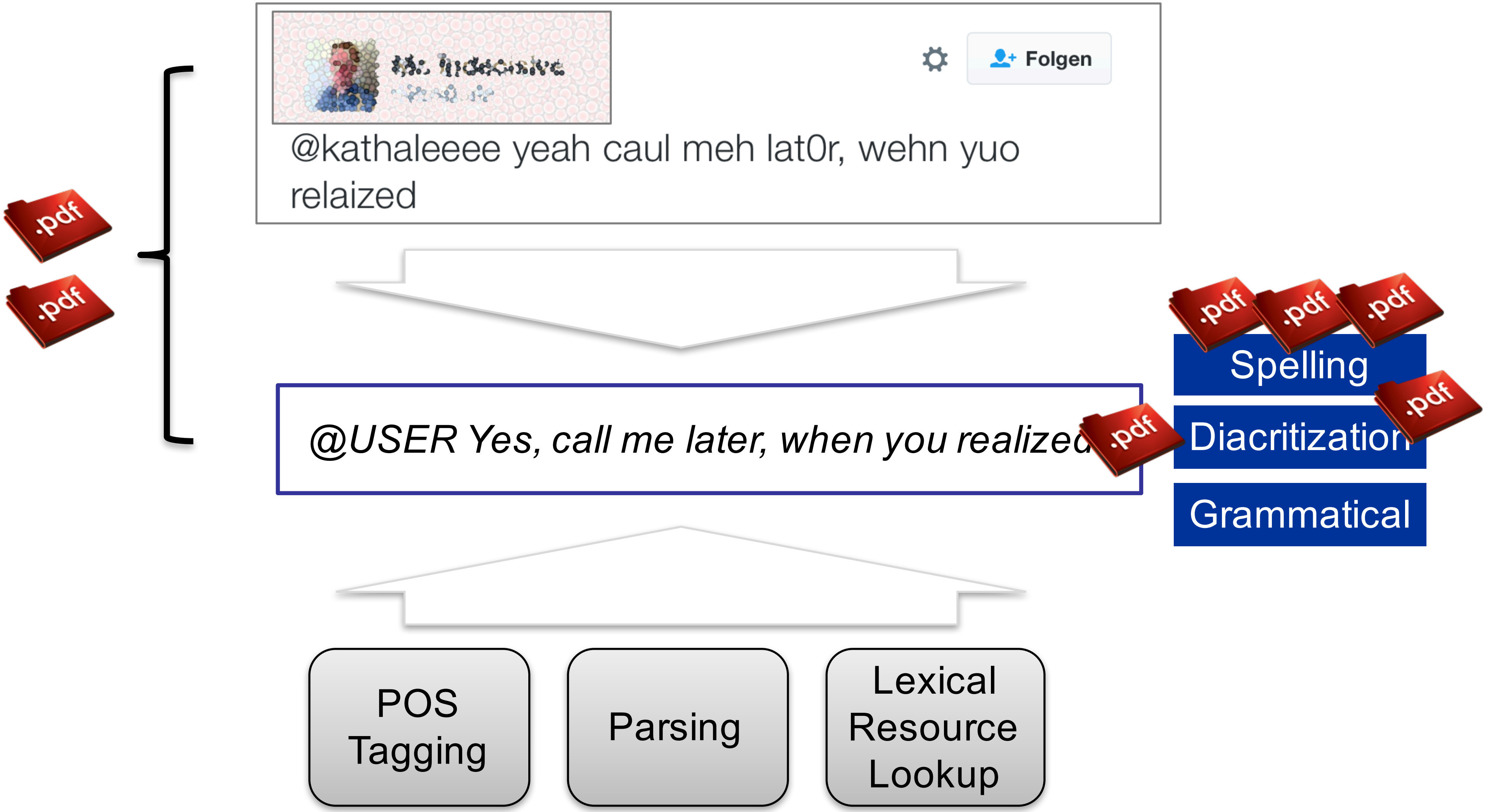
Social Media – Adapting Tools



Social Media – Adapting Data

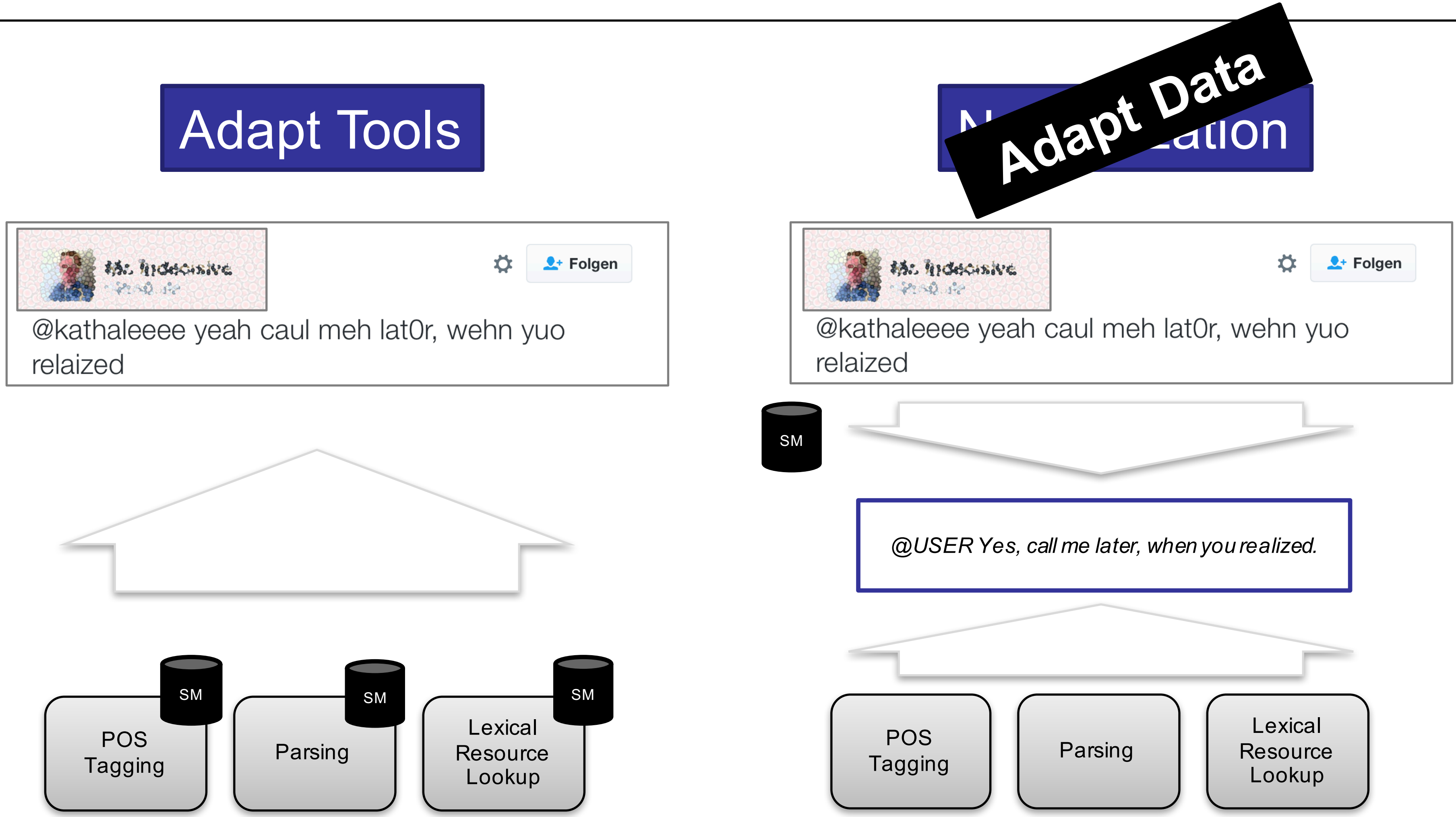


Locating the Workshop Papers



Comparing the Paradigms

UNIVERSITÄT DUISBURG ESSEN
Open-Minded



Comparing the Paradigms

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

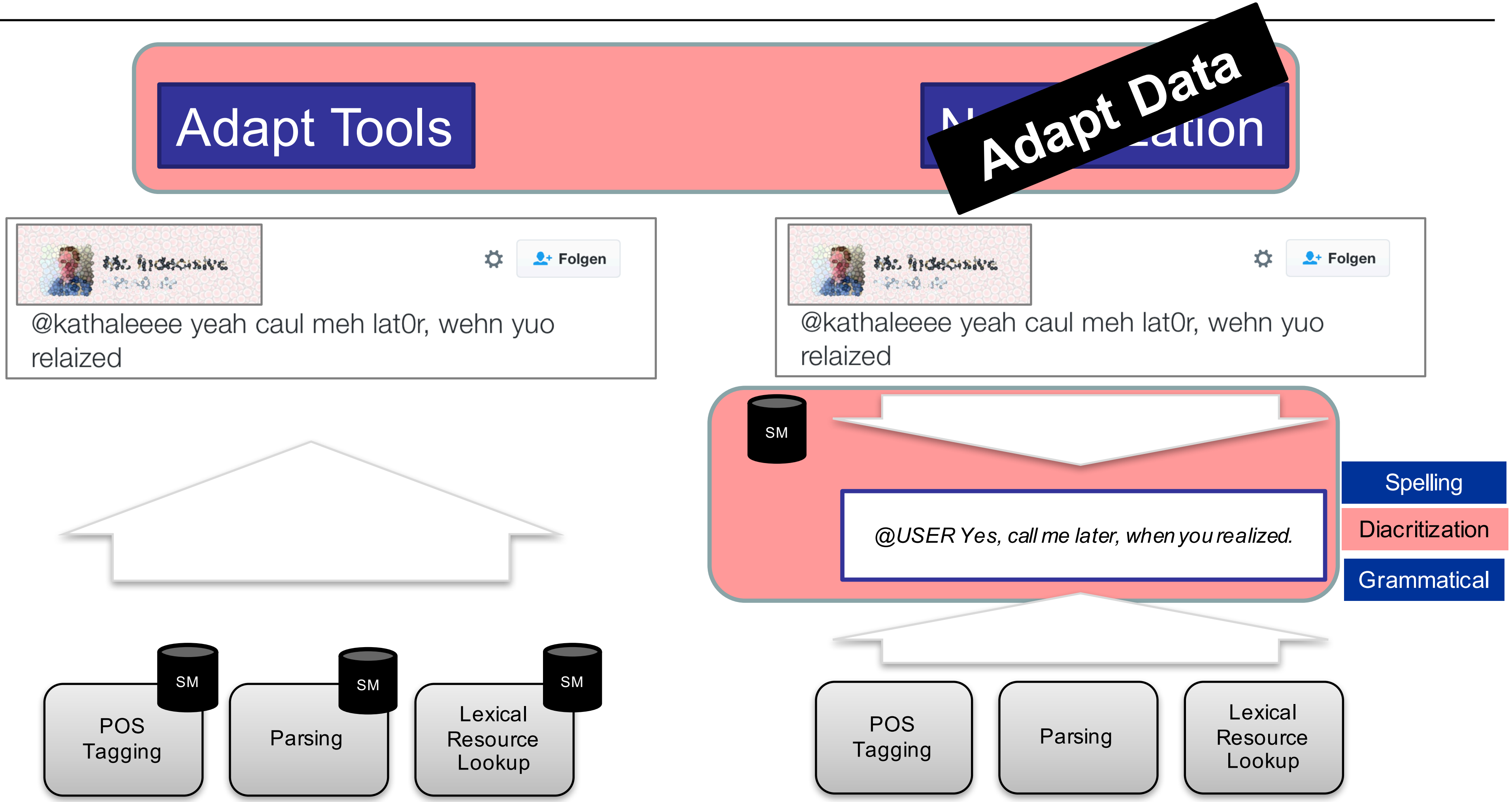
Adapt Tools

- Optimize task you care about
- Repeat for each task
- …

Adapt Data

- (possibly) improvements in many tasks
- Only once
- …

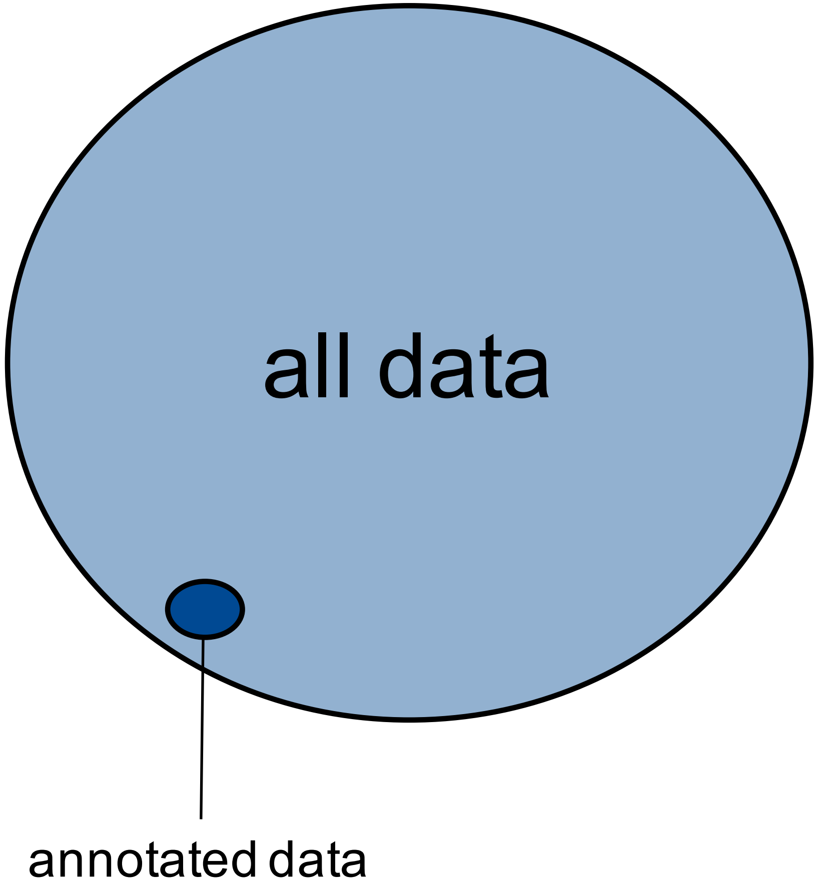
Talk Outline



POS Tagging

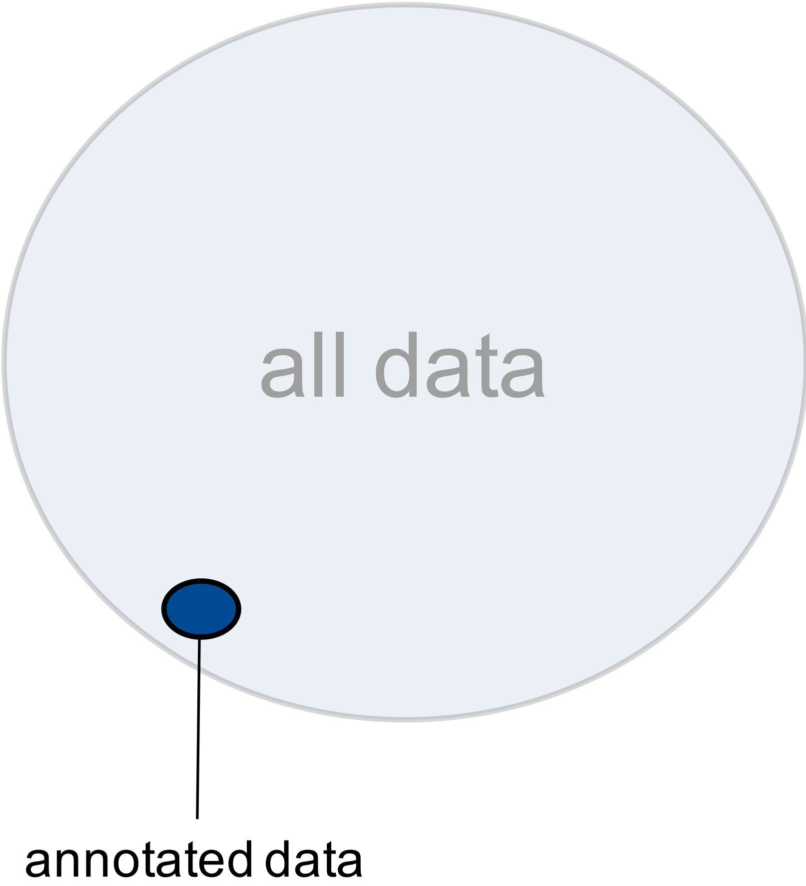
Supervised

Annotated Data



- Can only manually annotate/normalize tiny subset
- random sample is unlikely to contain rare phenomena

Annotated Data

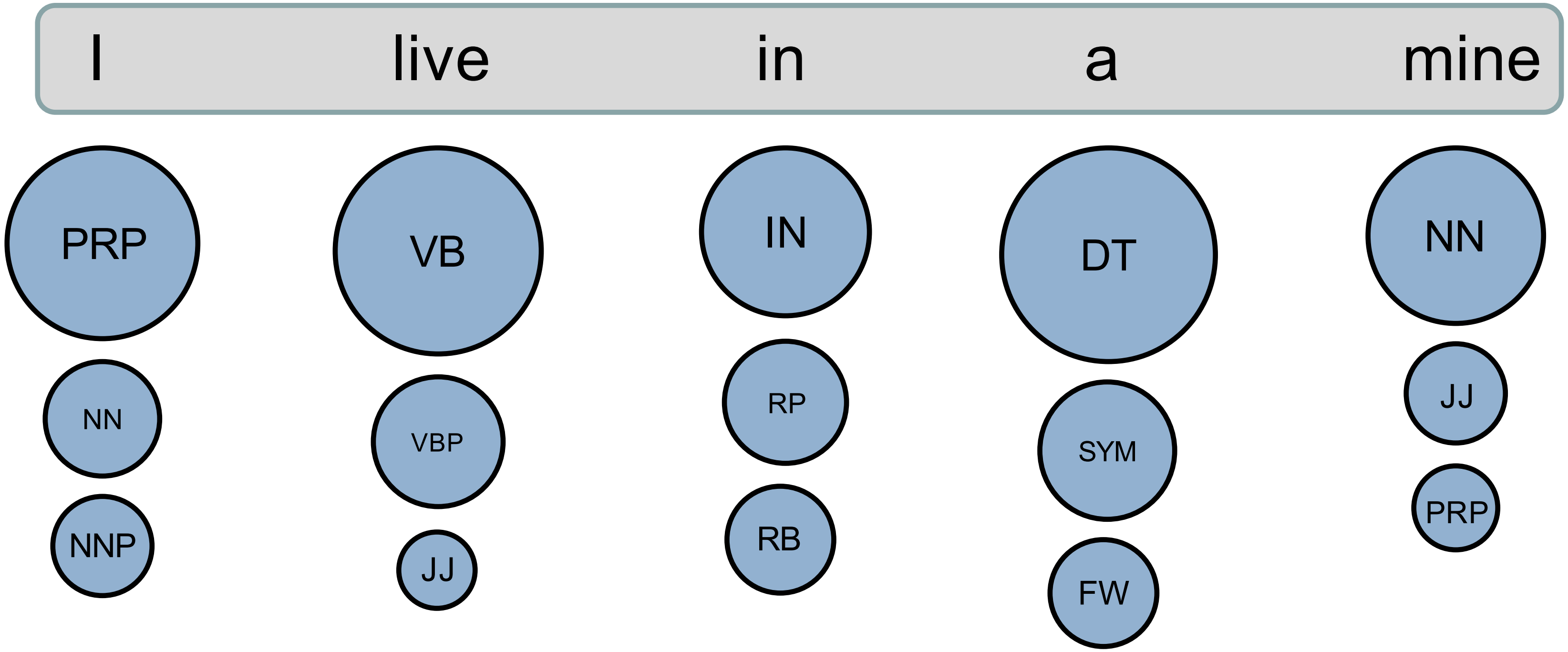


- Can only manually annotate/normalize tiny subset
- random sample is unlikely to contain rare phenomena

Long Tail

Rank	Tag	Frequency
1	NN	632
2	$.	351
…	…	…
50	PIDAT	5
51	PTKA	5
52	PWAT	4
53	TRUNC	4
54	VAPPER	4
55	VVIZU	3
56	FM	3
57	DM	2
58	APZR	2
59	KOUI	2
60	VMPPER	1
61	ADVART	1
62	KOUSPPER	1
63	PPERPPER	1
64	ONO	1

Statistical Information
PoS Distribution



Many OOV Words in Social Media

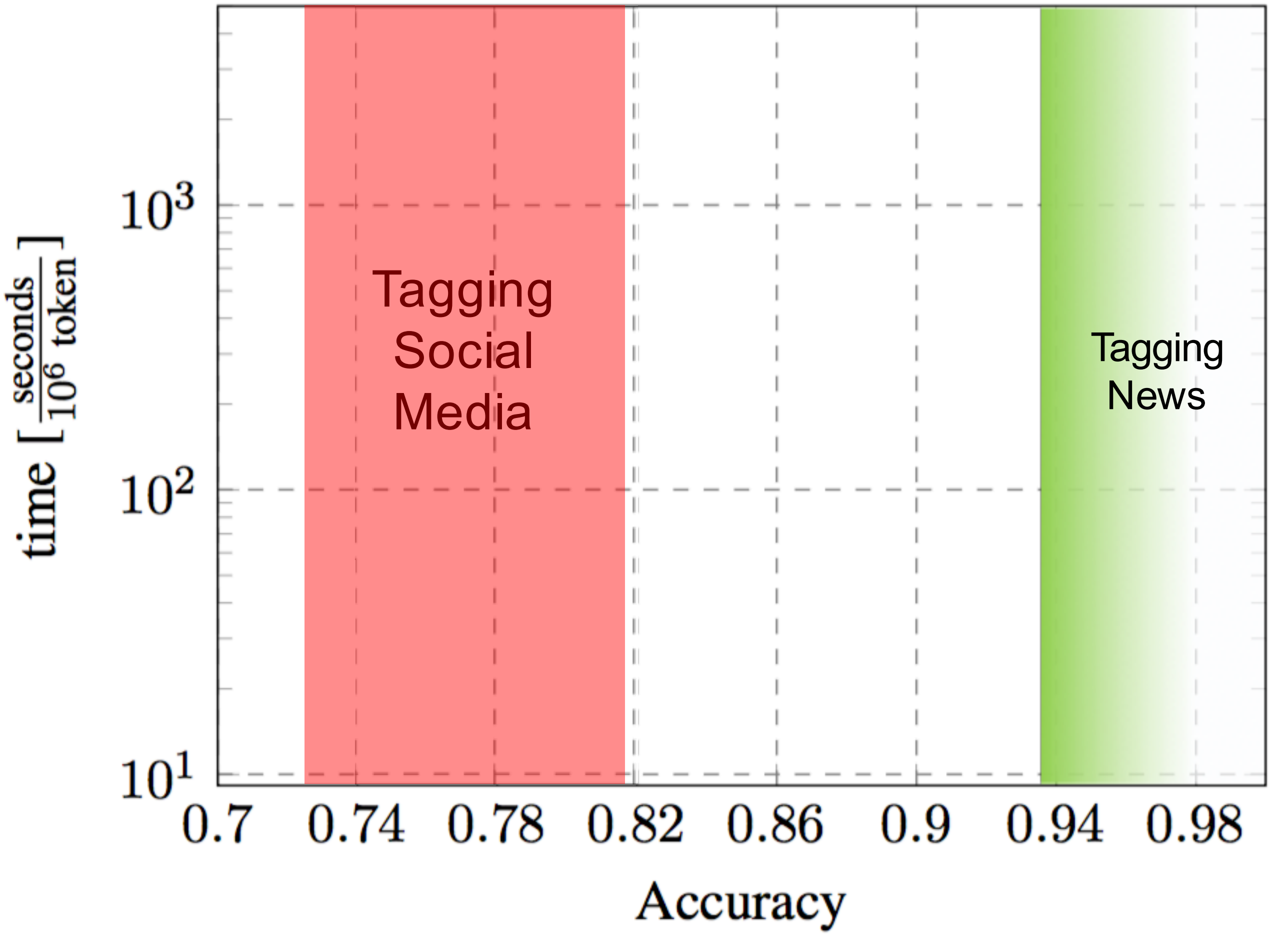
@kathaleeee yeah caul meh lat0r, wehn yuo relaized

What to do?

Use an available tagger

Tool	Language	Trained on	Modelname	Tagset	Domain	Abbr.
Ark	en	Owoputi	default	Gimpel	social	A-1
		Irc	irc	PTB-NPS	social	A-2
		Ritter	ritter	PTB-RIT	social	A-3
ClearNLP	en	Medical text	mayo	PTB	clinical	C-1
		OntoNotes	ontonotes	PTB	news	C-2
Hepple	en	*rule-based*		PTB	-	Hepple
HunPos	en	WSJ	wsj	PTB	news	Hun
	de	Tiger	tiger	STTS	news	
Mate	en	CoNLL2009	conll2009	PTB	mixed	Mate
	de	Tiger	tiger	STTS	news	
Lbj	en	WSJ	-	PTB	news	Lbj
OpenNLP	en	*unknown*	maxent	PTB	*unknown*	O-1
		unknown	perceptron	PTB	*unknown*	O-2
	de	Tiger	maxent	STTS	news	O-3
		Tiger	perceptron	STTS	news	O-4
Stanford	en	WSJ	bidirectional-distsim	PTB	news	St-1
		WSJ	caseless-left3w.-distsim	PTB	news	St-2
		unknown	fast	PTB	*unknown*	St-3
		Twitter/WSJ	twitter-fast	PTB-RIT	mixed	St-4
		Twitter/WSJ	twitter	PTB-RIT	mixed	St-5
		WSJ	wsj-0-18-caseless-left3w.-distsim	PTB	news	St-6
	de	Negra	dewac	STTS	news	St-7
		unknown	fast-caseless	STTS	news	St-8
		Negra	fast	STTS	news	St-9
		Negra	hgc	STTS	news	St-10
TreeTagger	en	*unknown*	le	PTB-TT	news	Tree
	de	*unknown*	le	STTS	news	

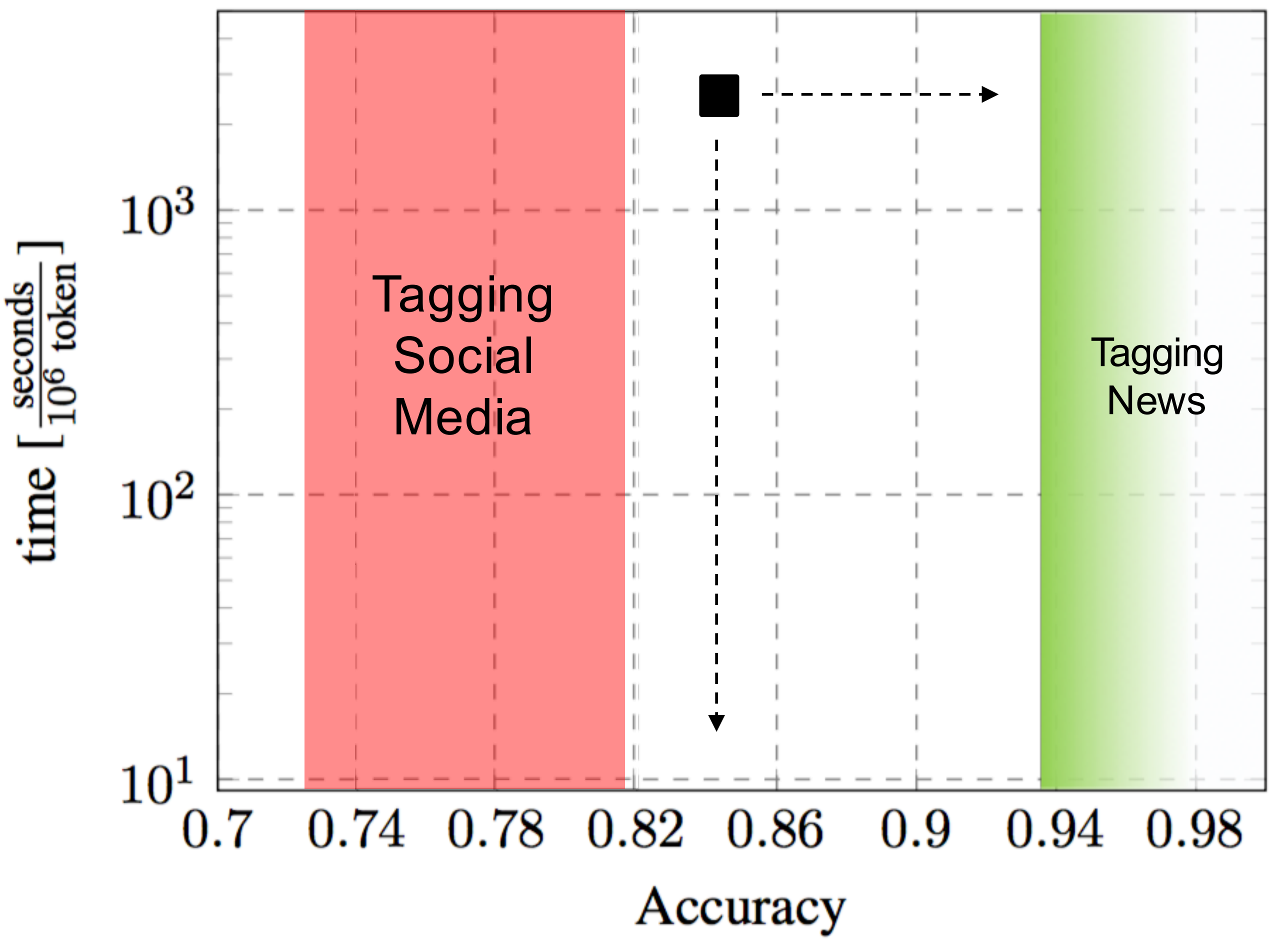
State-of-the-Art: Social Media vs. News



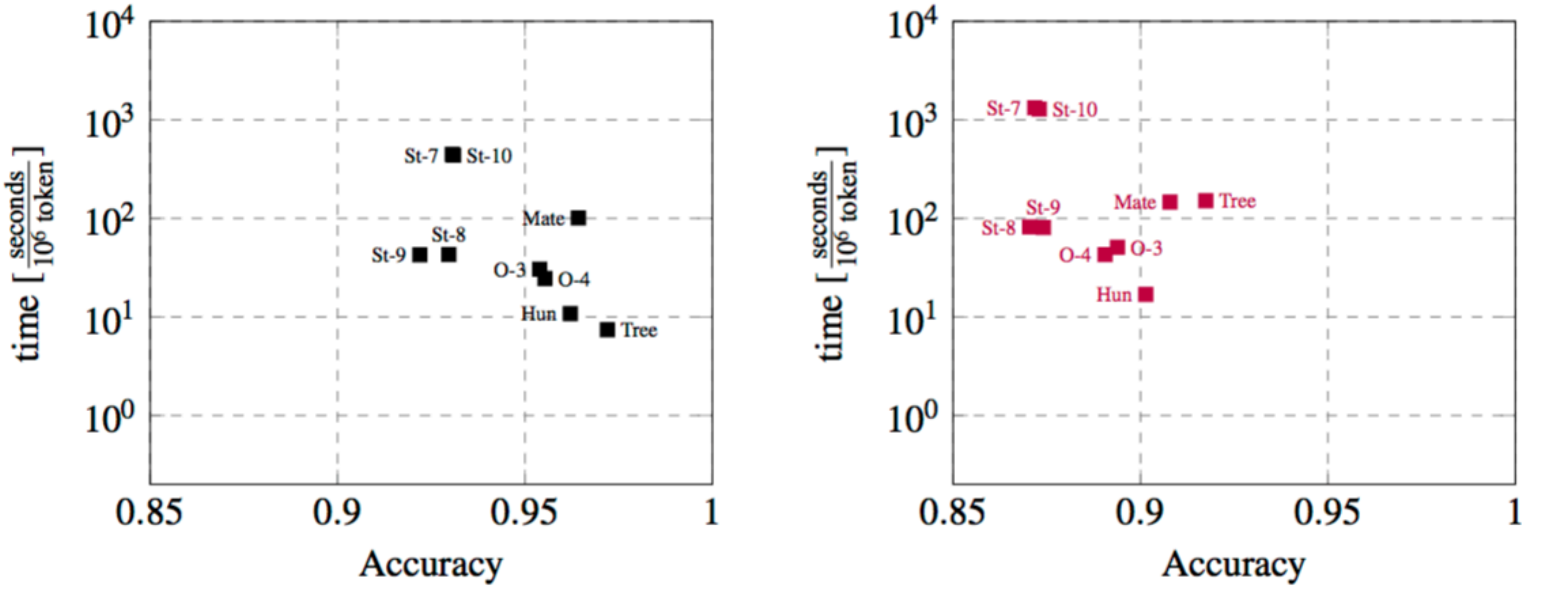
Improved Tagging on Social Media

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

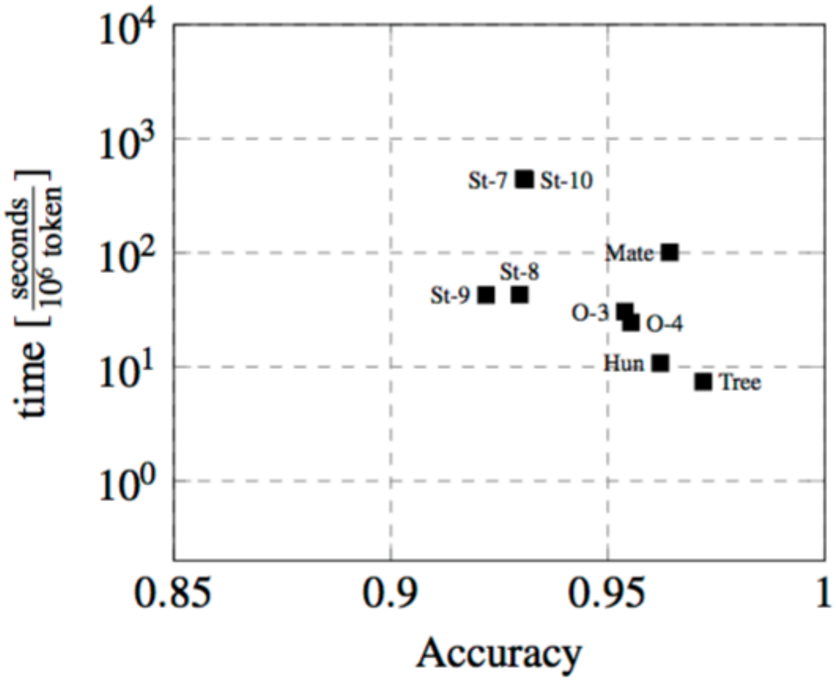


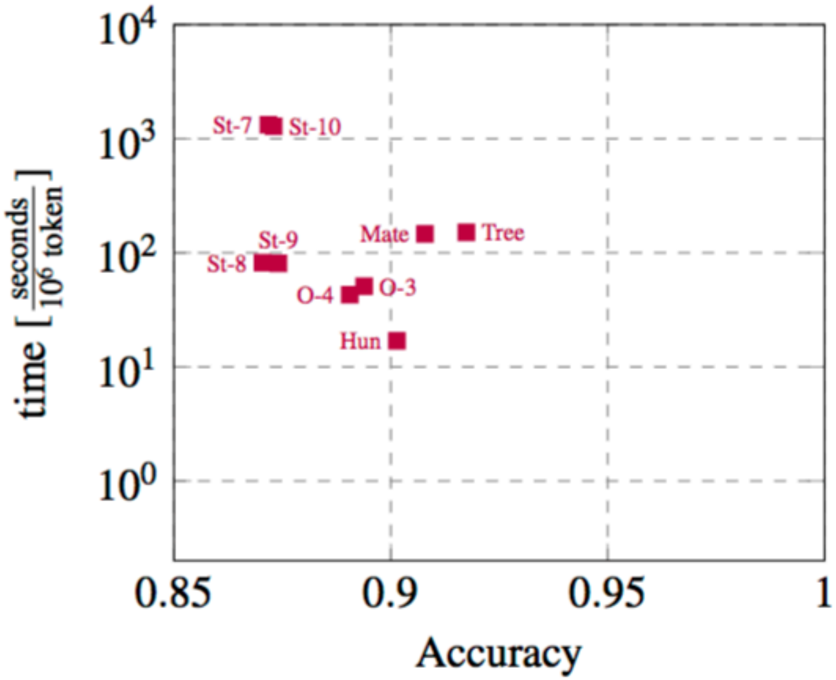
German: Written vs Social



(a) Written (b) Social

German: Written vs Social



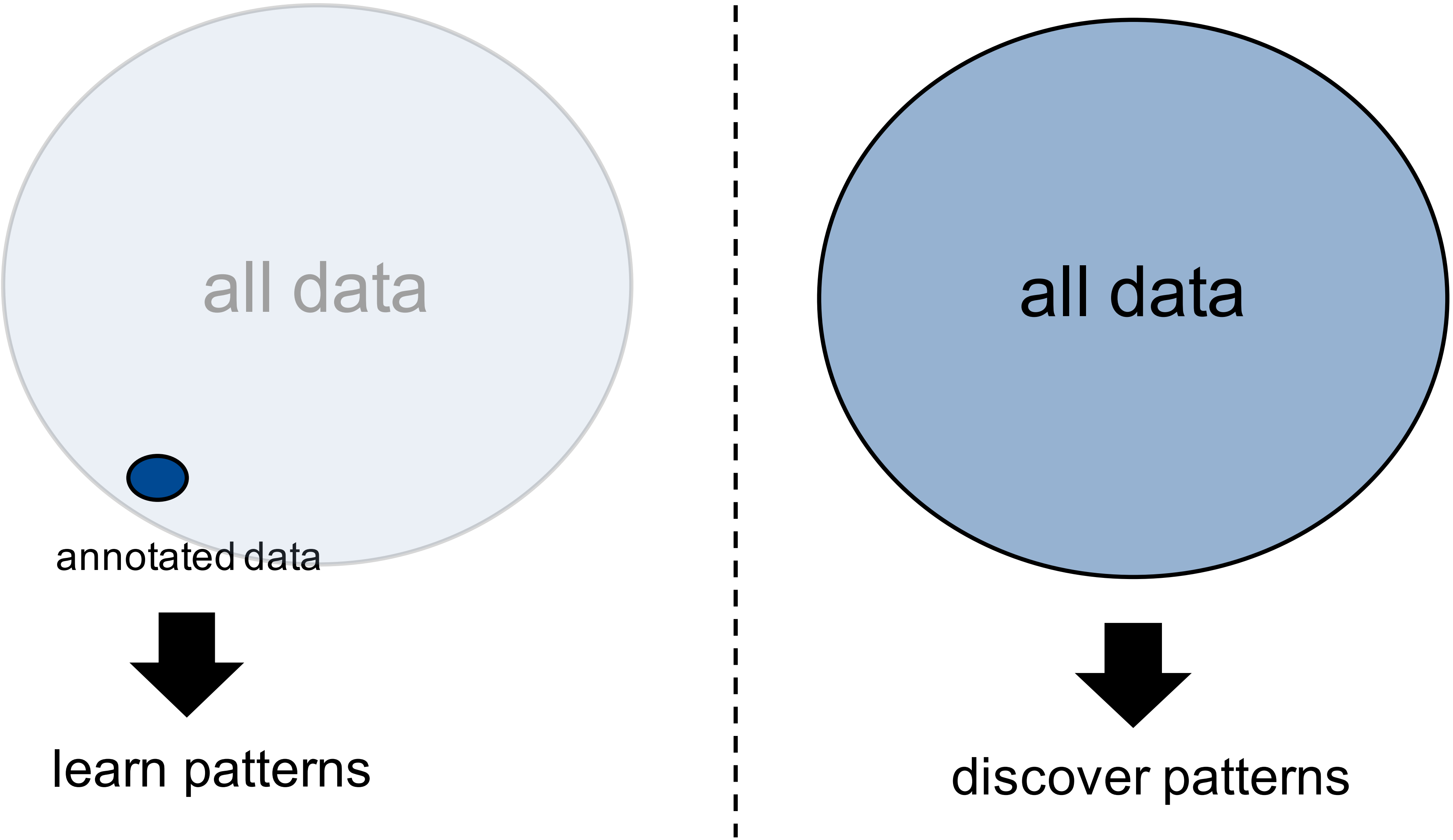


Let’s normalize!

POS Tagging

Unsupervised

Supervised vs. Unsupervised



Distributional Hypothesis

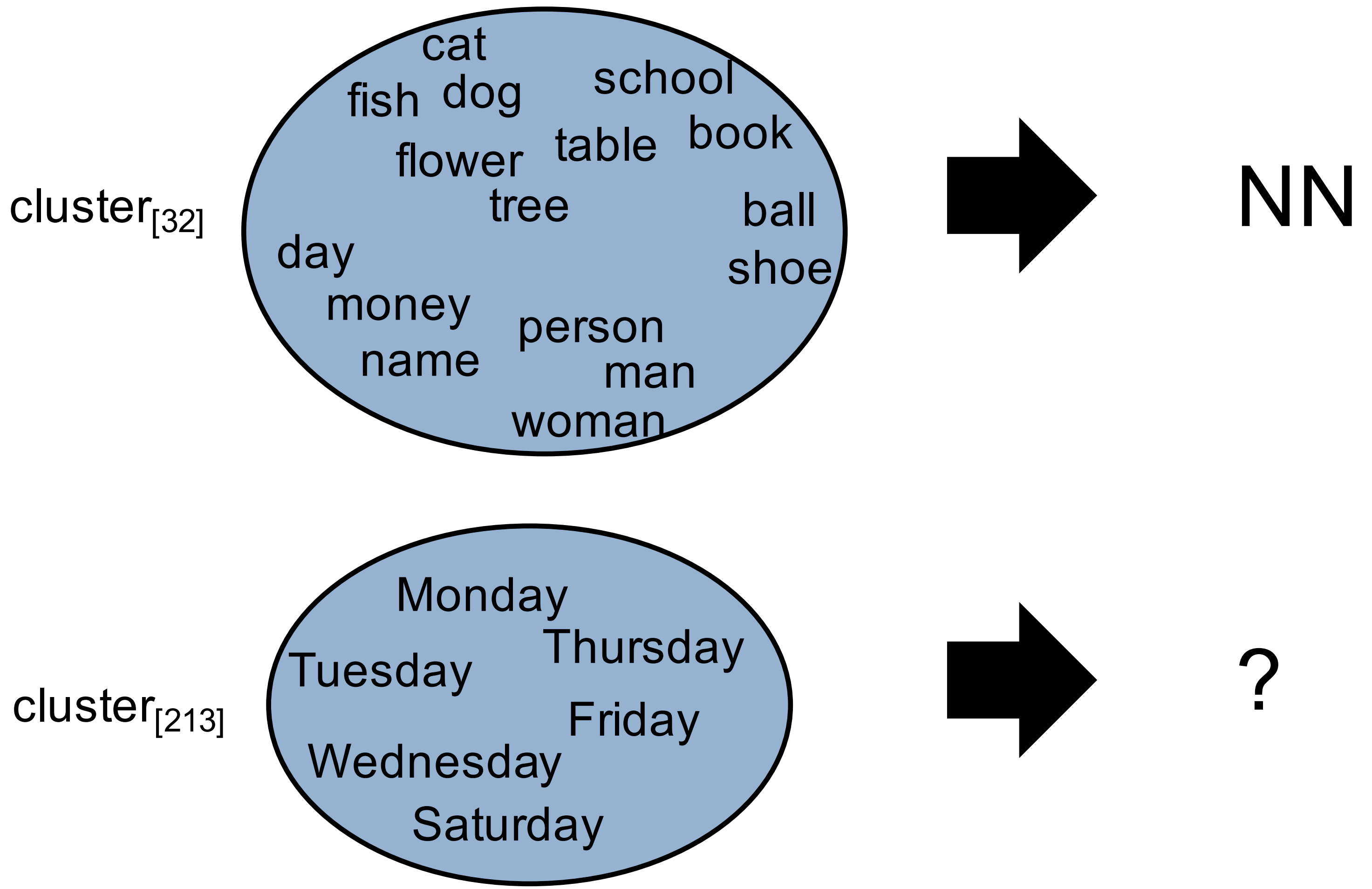
- A bottle of **tezguino** is on the table.
- Everybody likes **tezguino**.
- **Tezguino** makes you drunk.
- We make **tezguino** out of corn.

- *"a word is characterized by the company it keeps"* (Firth, 1957)

- in a cluster with

Wine
Beer
Vodka
Whiskey
Rum

Clusters vs. Word Classes



Problem 1: Sparse POS classes

Rank	Tag	Frequency
1	NN	632
2	$.	351
…	…	…
50	PIDAT	5
51	PTKA	5
52	PWAT	4
53	TRUNC	4
54	VAPPER	4
55	VVIZU	3
56	FM	3
57	DM	2
58	APZR	2
59	KOUI	2
60	VMPPER	1
61	ADVART	1
62	KOUSPPER	1
63	PPERPPER	1
64	ONO	1

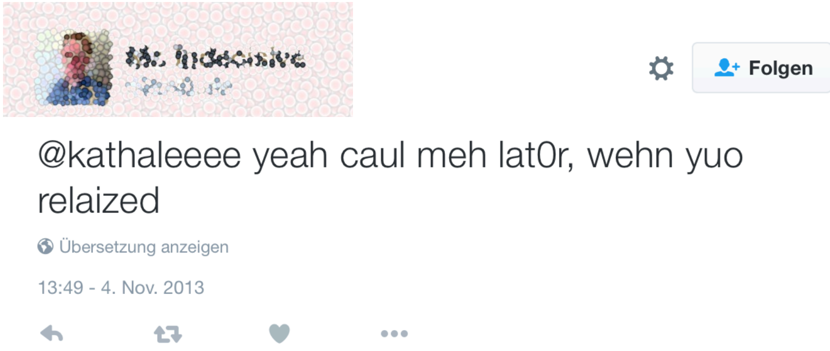
Problem 2: OOV Word

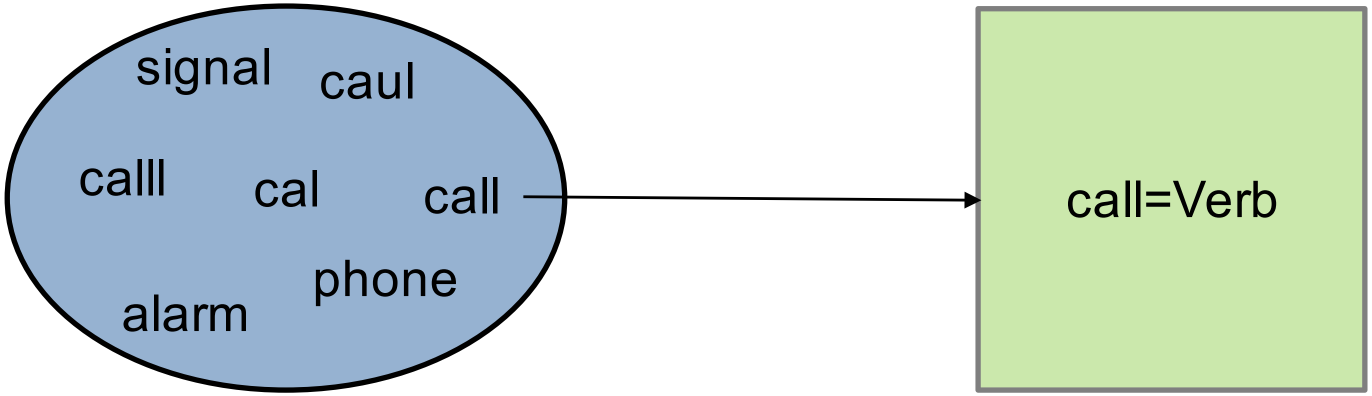
What to do?

Improving Supervised with Unsupervised

cluster$_{[513]}$

If you know one word's PoS …





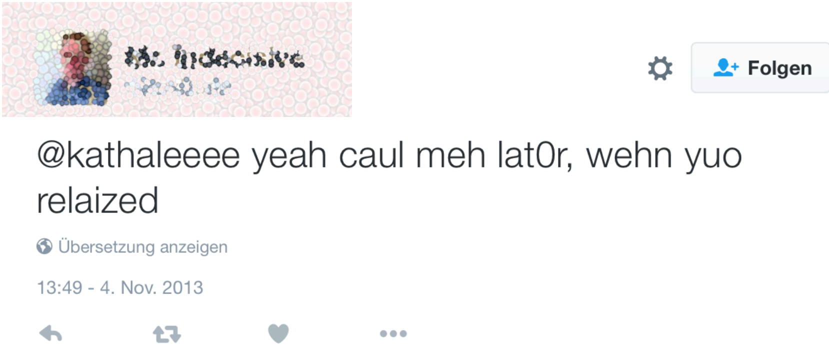
cluster$_{[513]}$

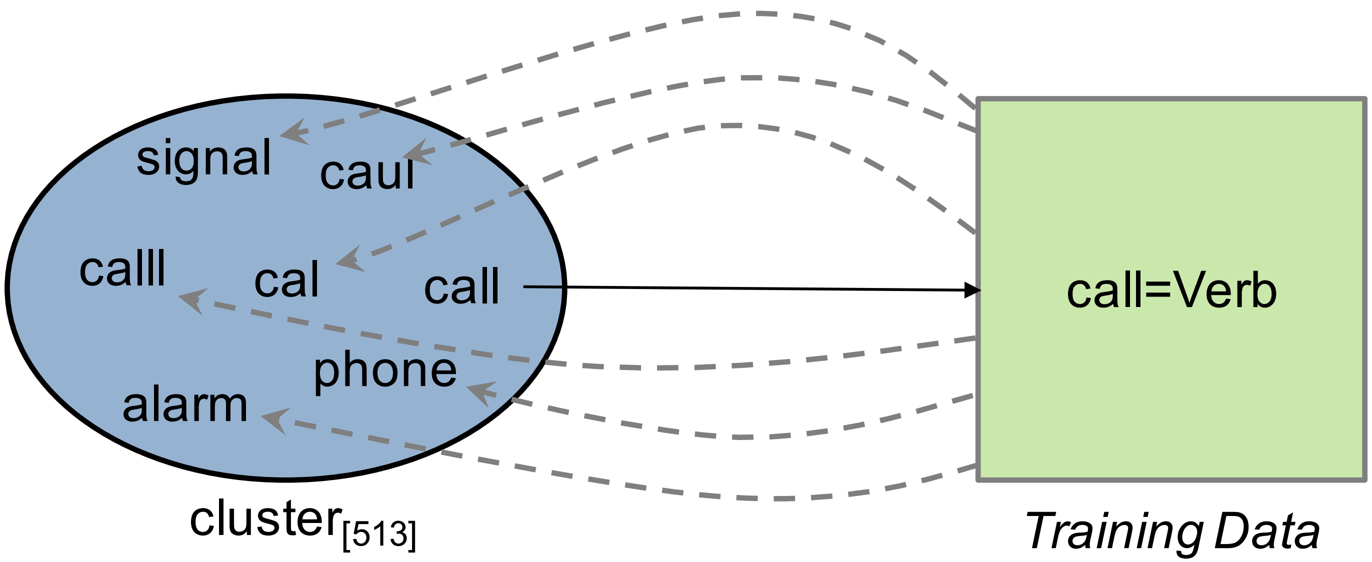
Training Data

…you know them all

UNIVERSITÄT DUISBURG ESSEN

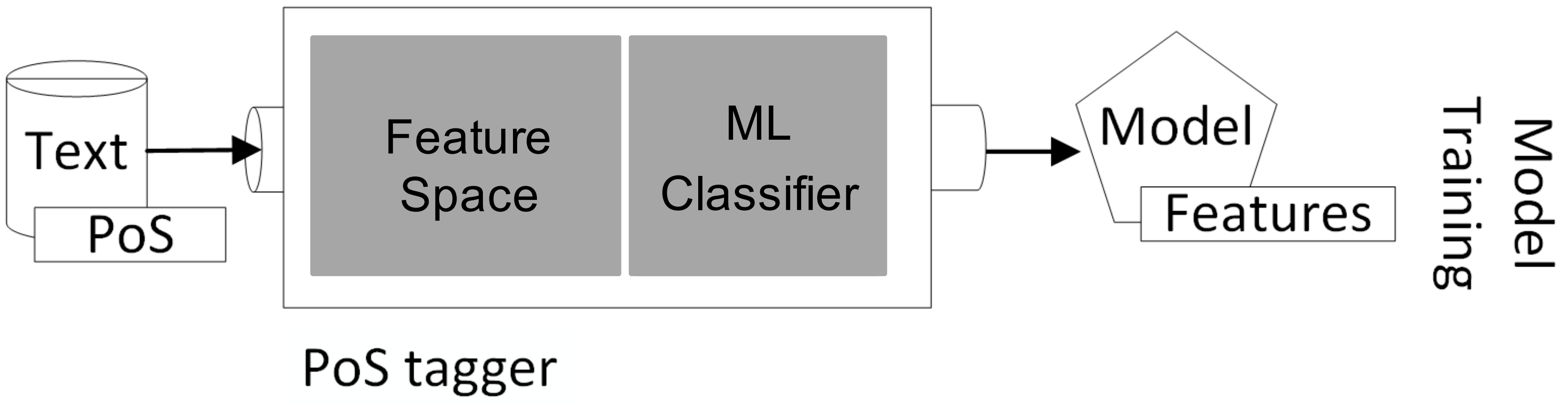
Open-Minded





PoS tagger are black boxes

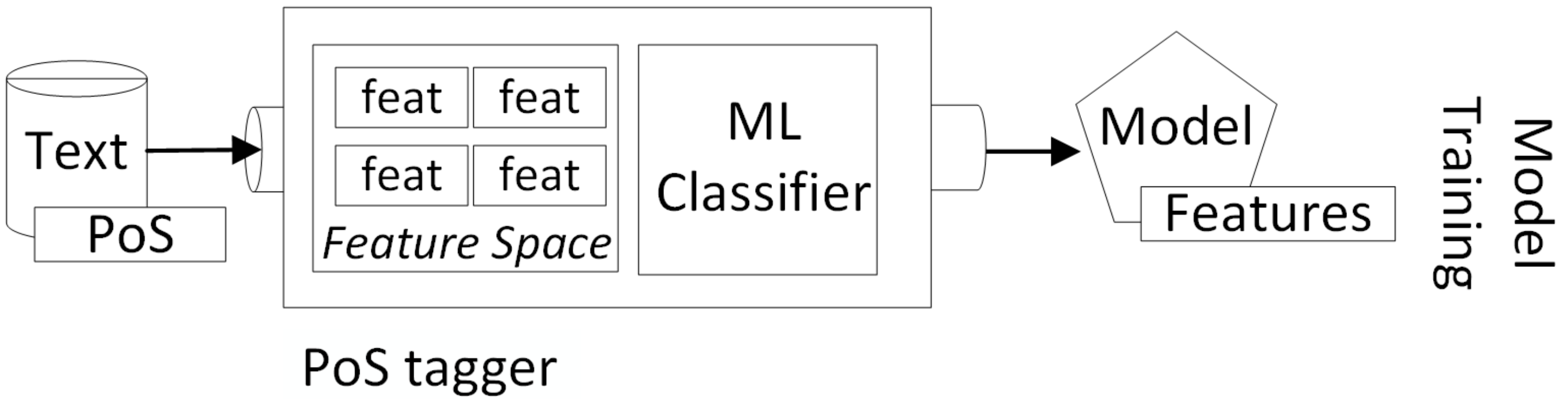
- Feature space is fixed
- New resources cannot be easily integrated



FlexTag: A Highly Flexible PoS Tagger

- Feature space is fixed
- New resources cannot be easily integrated

- *“FlexTag: A feature space exposing PoS tagger” (LREC, 2016)*
- Easy to experiment with new features while being still easy to use



EmpiriST PoS Tagging Shared Task

GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / Social Media

- German dataset
- CMC and Web data

Ranks	Teams	Overall-Acc
1	UdS	90.44
2	LTL-UDE	89.09
3	AIPHES	88.75
4	bot.zen	88.03
5	COW	84.86
6	$WAGMOB	84.64

Results by Genre

	CMC		Web		∅	
	Generic	ST-specific	Generic	ST-specific	Generic	ST-specific
TreeTagger	73.8	77.3	91.6	91.8	84.2	84.6

Results by Genre

	CMC		Web		∅	
	Generic	ST-specific	Generic	ST-specific	Generic	ST-specific
TreeTagger	73.8	77.3	91.6	91.8	84.2	84.6
EmpiriST	72.2	73.4	75.5	76.3	73.9	74.9

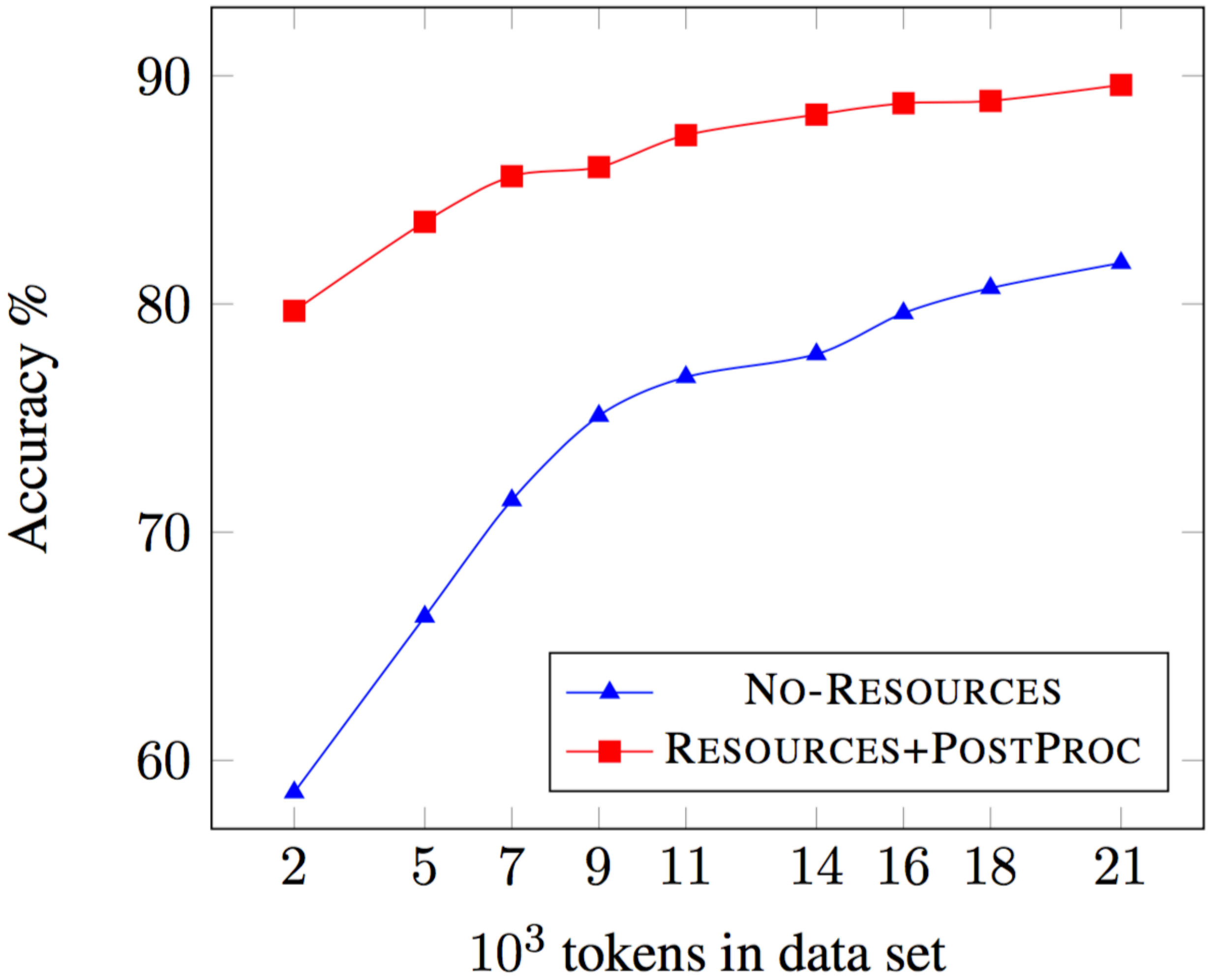
Results by Genre

	CMC		Web		∅	
	Generic	ST-specific	Generic	ST-specific	Generic	ST-specific
TreeTagger	73.8	77.3	91.6	91.8	84.2	84.6
EmpiriST	72.2	73.4	75.5	76.3	73.9	74.9
+Tiger	79.6	80.6	88.8	88.9	84.2	84.8
+Tiger+Brown	84.4	85.2	90.8	90.6	87.6	87.9
+Tiger+MorphLex	81.1	81.5	90.6	90.8	85.9	86.2
+Tiger+PosDict	82.4	83.8	91.0	91.4	86.7	87.6

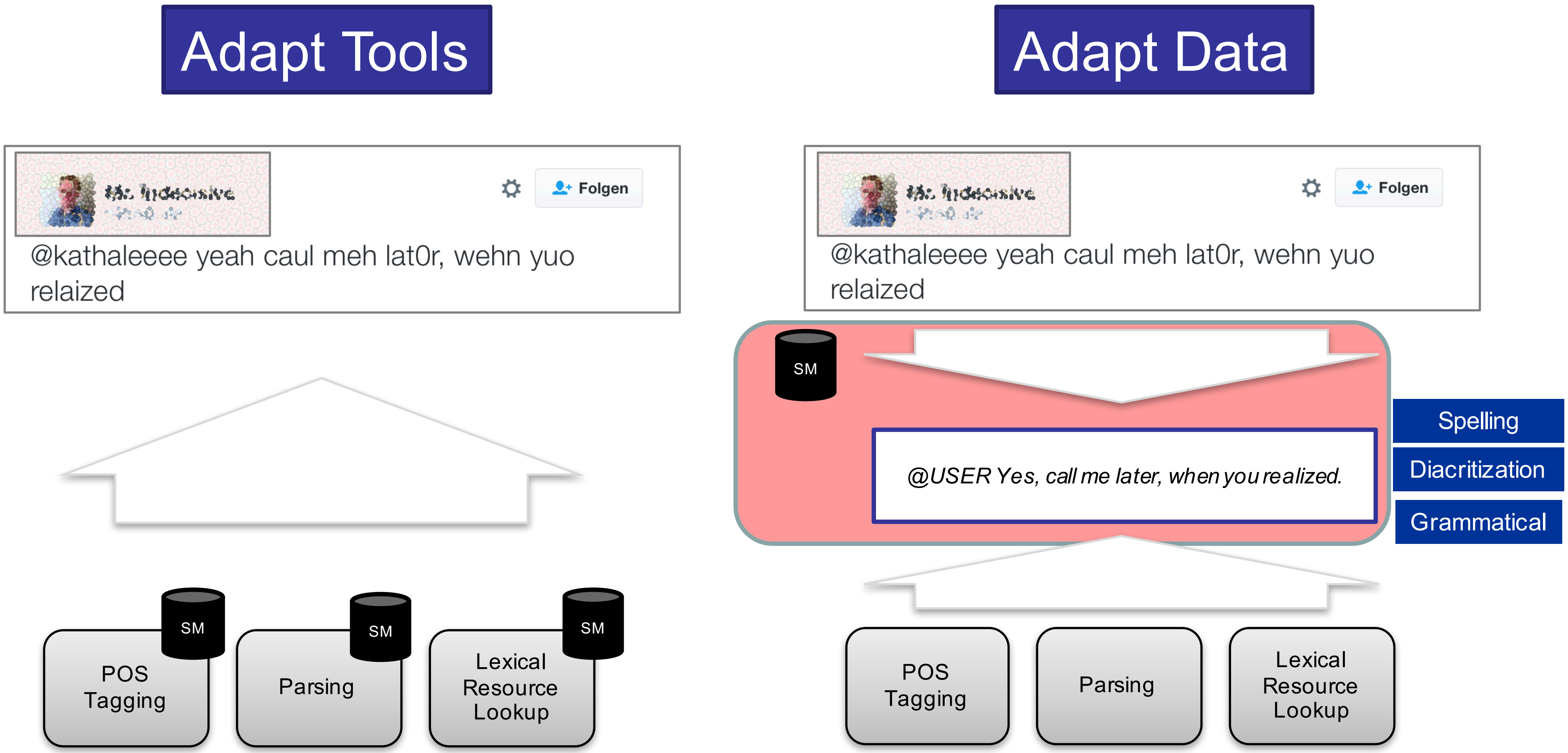
Results by Genre

	CMC		Web		∅	
	Generic	ST-specific	Generic	ST-specific	Generic	ST-specific
TreeTagger	73.8	77.3	91.6	91.8	84.2	84.6
EmpiriST	72.2	73.4	75.5	76.3	73.9	74.9
+Tiger	79.6	80.6	88.8	88.9	84.2	84.8
+Tiger+Brown	84.4	85.2	90.8	90.6	87.6	87.9
+Tiger+MorphLex	81.1	81.5	90.6	90.8	85.9	86.2
+Tiger+PosDict	82.4	83.8	91.0	91.4	86.7	87.6
All resources	85.6	86.1	92.0	92.1	88.8	89.1

Learning Curve



Talk Outline



Practical Limits

SemEval 2016 Task on Stance Detection

Task: Classify whether the author of a tweet is
(a) in favor
(b) against or
(c) if neither inference is likely

Five target domains:
- *Atheism*
- *Abortion*
- *Feminism*
- *Hillary Clinton*
- *Climate change*

Examples

Target: Atheism

>>**Could all those who believe in god please leave. The meeting will now continue for the grown ups only.** <<

Target: Hillary Clinton

>>**They should just make the GOP primaries a reality game show called "Who Wants To Get Beat Up By A Girl?"**<<

Example: Classifying Stance

Target: Atheism

FAVOR

god, #freethinker, religion, evidence
Christianity, superstitions, bullshit,
impossible, evidence-based,
Atheist, #FuckReligion, …

AGAINST

God, #God, Jesus, islam, Him,
Amen, Holy, Christ, Lord, Allah,
#spirituality, prayer, Quran, …

god

God

Theoretical Limits

How to normalize?

Normalize Learner Data [Lüdeling et al.]

Normalization always wrt. a Target Hypothesis

- Definition of normalization goal
- Requires interpretation
- Influenced by research question
- Does not encode "truth" or "correct" usage

Error

- Difference between a learner utterance and a target hypothesis

- Lüdeling, Anke, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter (2008). "Das Lernerkorpus Falko". In: *Deutsch als Fremdsprache* 45.2, pp. 67–73.
- Reznicek, Marc, Anke Lüdeling, and Hagen Hirschmann (2013). "Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture". In: *Automatic Treatment and Analysis of Learner Corpus Data*. Ed. by Ana Díaz-Negrillo, Nicholas Ballier, and Paul Thompson. Amsterdam: John Benjamins, pp. 101–124.
- Lüdeling, Anke and Hagen Hirschmann (2015). "Error Annotation". In: *The Cambridge Handbook of Learner Corpus Research*. Ed. by Sylviane Granger, Gaetanelle Gilquin, and Fanny Meunier. Cambridge: Cambridge University Press.

Example

bevor	man	überhaupt	anfangen	kann,	sich	neues	Wissen	zu erlernen
before	one	even	start	can	REFL	new	knowledge	to learn

'before one can even start to acquire new knowledge'

Example – Possible Corrections

bevor	man	überhaupt	anfangen	kann,	sich	neues	Wissen	zu erlernen
bevor	man	überhaupt	anfangen	kann,		neues	Wissen	zu erlernen

→ argument structure error

Example – Possible Corrections

UNIVERSITÄT DUISBURG ESSEN

Open-Minded

bevor man überhaupt anfangen kann, sich neues Wissen zu erlernen

bevor man überhaupt anfangen kann, neues Wissen zu erlernen

→ argument structure error

bevor man überhaupt anfangen kann, sich neues Wissen anzueignen

→ lexical error

Example – Possible Corrections

bevor man überhaupt anfangen kann, sich neues Wissen zu erlernen

---

bevor man überhaupt anfangen kann, neues Wissen zu erlernen
→ argument structure error

bevor man überhaupt anfangen kann, sich neues Wissen anzueignen
→ lexical error

bevor man überhaupt anfangen kann, neues Wissen zu erwerben
→ lexical error and argument structure error

Social Media

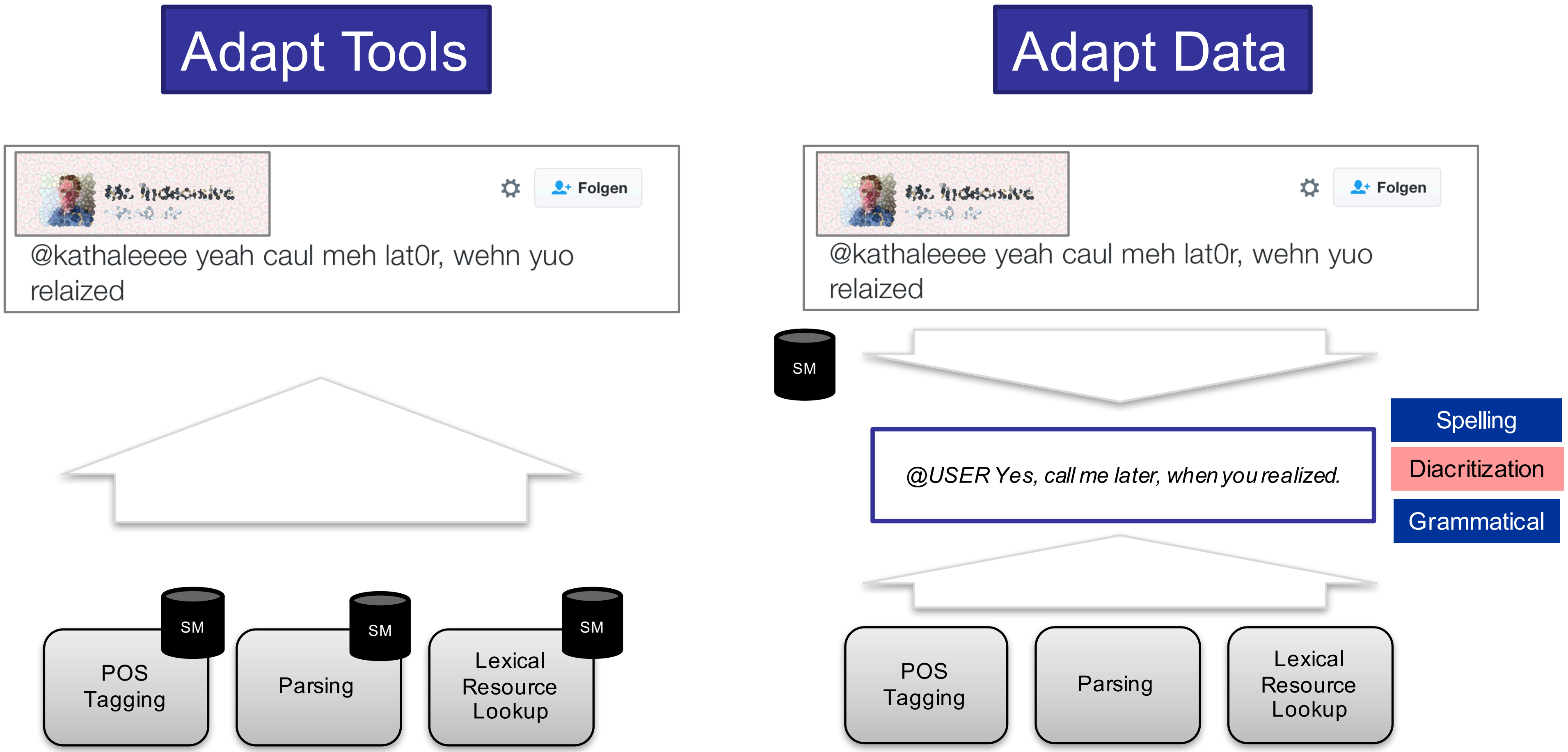
- Usually (implicitly) follows a spelling based target hypothesis

TH1: @kathaleeee Yeah, call me later, when you realized.
TH2: @USER Yes, call me later, when you have decided.

Talk Outline

UNIVERSITÄT DUISBURG ESSEN

Open-Minded



Diacritization = Normalization?

- All normalization is interpretation
- Unambiguous cases are trivial → so all normalization is disambiguation
- Diacritization is actually word sense disambiguation

High Ambiguity

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

[e□l□m□]

علم	[elm]	Science	عِلْمْ	[eilm]
		Flag	عَلَمْ	[ealam]
		He knew	عَلِمَ	[ealima]
		It was known	عُلِمَ	[eulima]

Modern Standard Arabic (MSA)

Written without diacritics (that represent mainly short vowels)

- wld ls wrk n nglsh, bt nt s wll

Reasons:

- Tradition
- Takes time to write
- Hard to read

عِلْمْ عِلْمْ

Do you know these words?

معمر القذافي	Muammar Al-Gaddafi
طالب	Talib

Do you know these words?

معمر القذافي mEmr Alq*Afy

طالب TAlb

L1-specific Transliterations

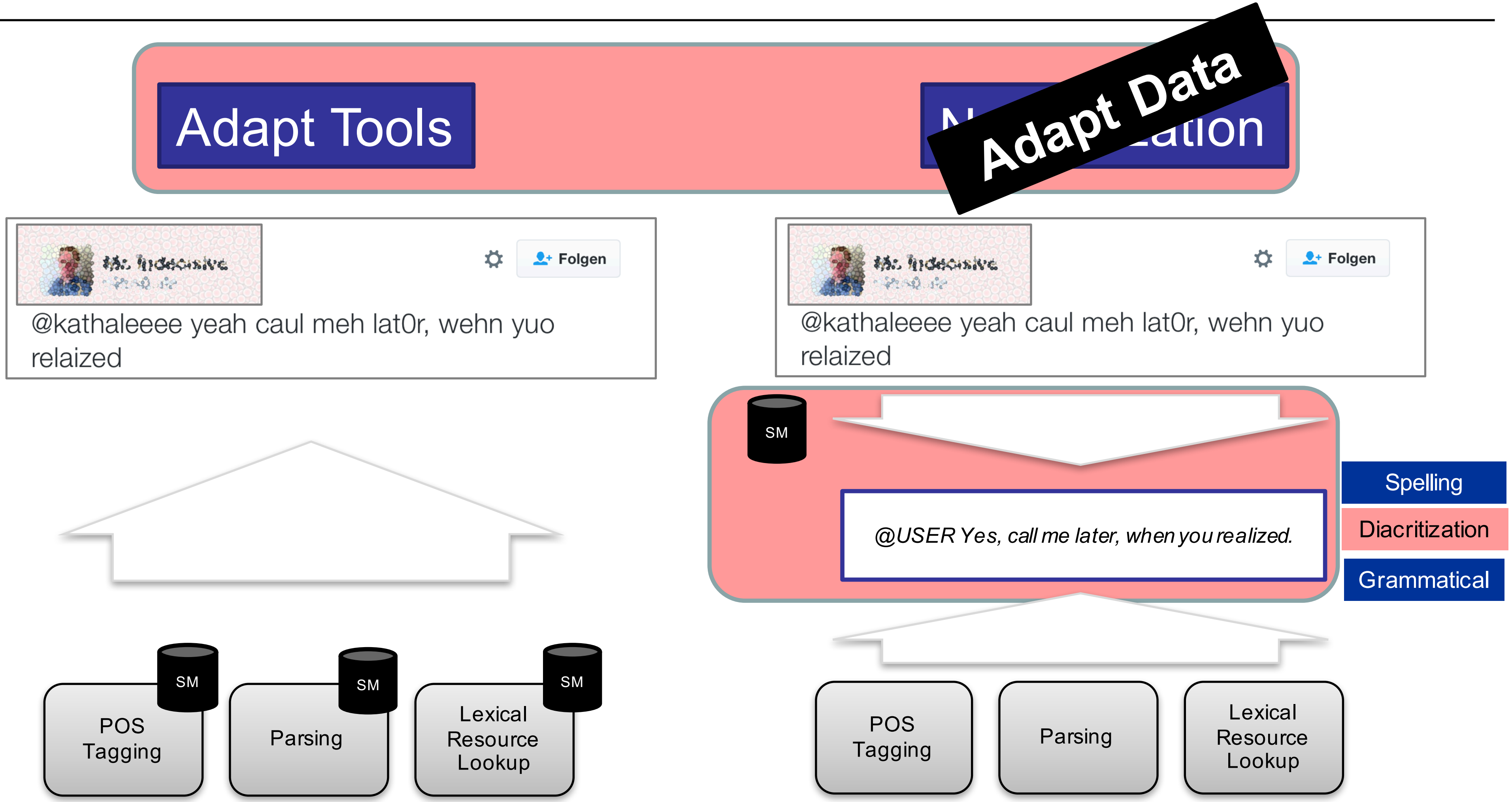
Muammar **Gaddafi**

- Gadafi
- Kaddafi
- Kathafi
- Kadhafi
- Kazafi
- Qathafi
- Qadafi
- Qadhafi
- Qazzafi
- …

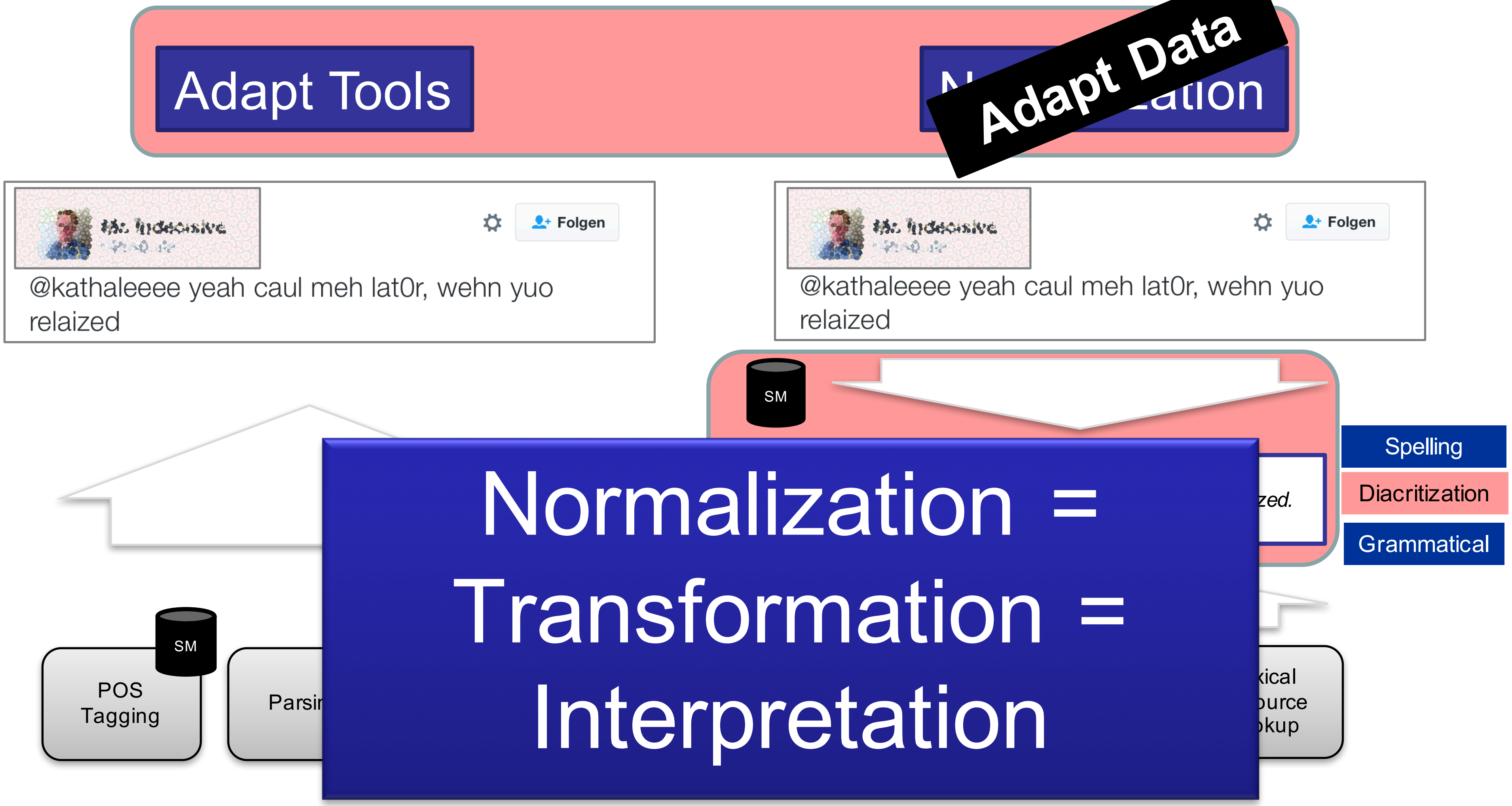
Czech	Kaddáfí
English	Gaddafi
German	Gaddafi
Italian	Gheddafi
Polish	Kaddafi
Spanish	Gadafi
Turkish	Kaddafi

Summary

Wrapping Up


Adapt Tools
Adapt Data
Folgen
@kathaleeee yeah caul meh lat0r, wehn yuo relaized
Folgen
@kathaleeee yeah caul meh lat0r, wehn yuo relaized
SM
@USER Yes, call me later, when you realized.
Spelling
Diacritization
Grammatical
SM
SM
SM
POS Tagging
Parsing
Lexical Resource Lookup
POS Tagging
Parsing
Lexical Resource Lookup

Wrapping Up



Adapt Tools
Adapt Data
@kathaleeee yeah caul meh lat0r, wehn yuo relaized
@kathaleeee yeah caul meh lat0r, wehn yuo relaized
SM
Spelling
Diacritization
Grammatical
SM
POS Tagging
Normalization = Transformation = Interpretation