Computational Linguistics & Resources from Social Media

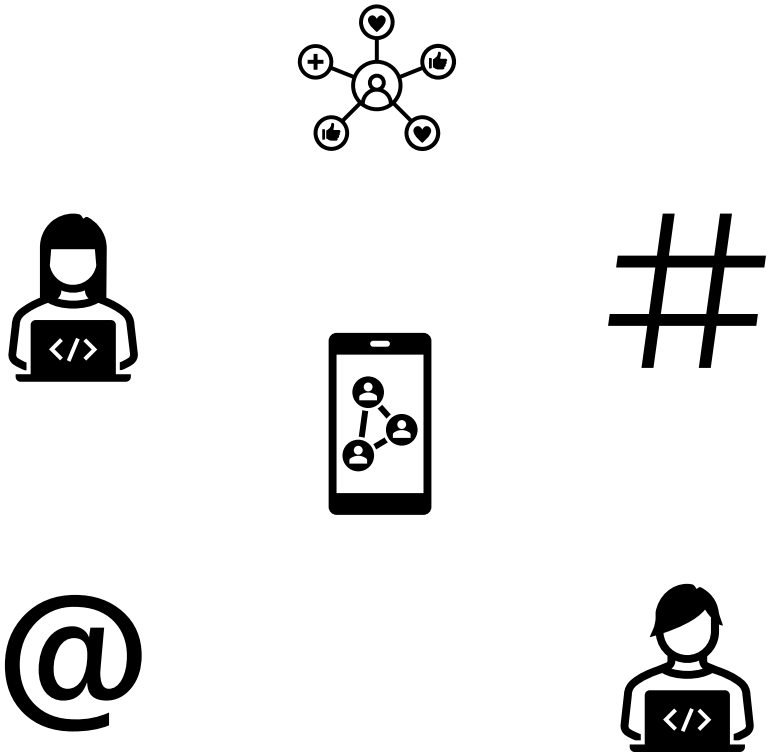
October 13th, 2022

Manuela Sanguinetti

Department of Mathematics and Computer Science, University of Cagliari

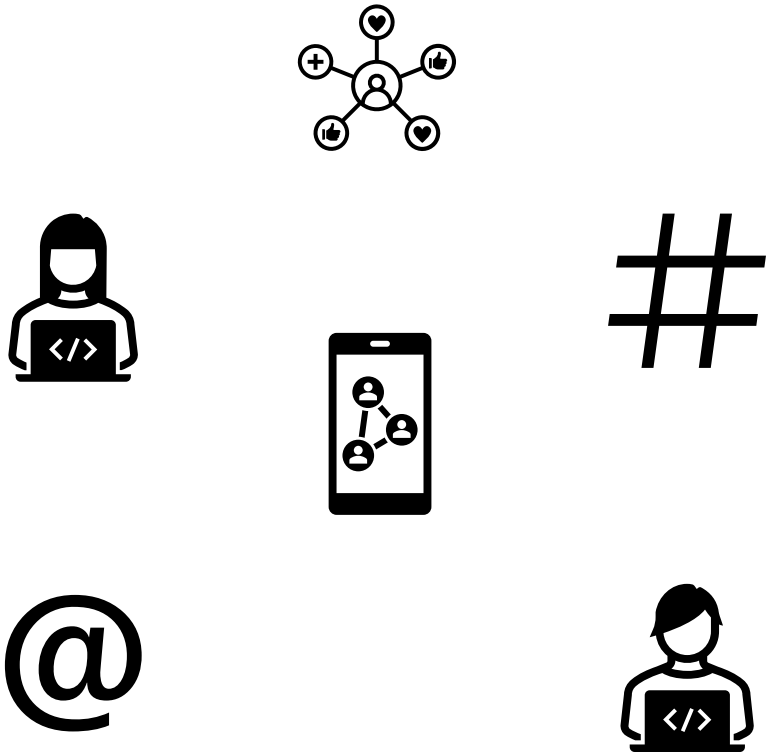
Overview

1. NLP & Social Media
2. Challenges & Issues in Resource Development
3. Guidelines for UGC Data
4. Hands-on Session



Overview

1. **NLP & Social Media**
2. Challenges & Issues in Resource Development
3. Guidelines for UGC Data
4. Hands-on Session



NLP & Social Media

- Social platforms as spaces for discussion on a variety of topics:

Provide new opportunities for influencing public opinion and for different voices to be heard

Raise questions about the legality or veracity of the content being broadcast

- **Natural Language Processing** as a powerful means to identify/analyze all these aspects
- CHALLENGE: the treatment (on different levels and for different purposes) of the so-called **User-Generated Content** (UGC), i.e., any Web content that comes in the form of images, videos, social media posts, reviews, etc.

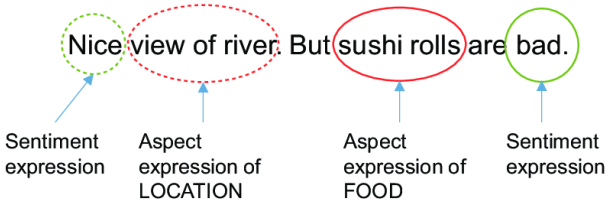
Examples

SENTIMENT

NLP & Social Media - Examples

SENTIMENT

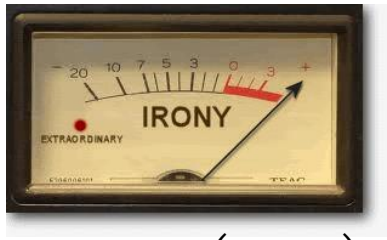
ASPECT



STANCE

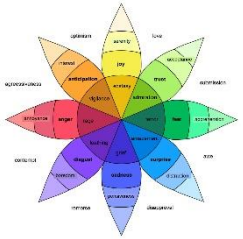
(source)

NLP & Social Media - Examples



(source)

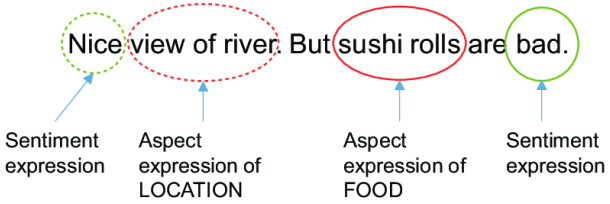
HUMOR/IRONY/SARCASM



EMOTION

SENTIMENT

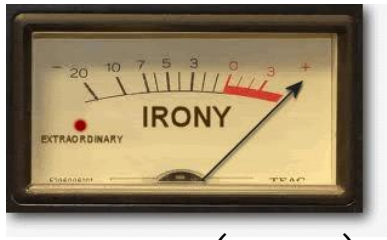
ASPECT



(source)

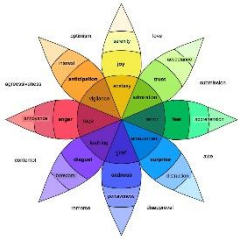
STANCE

NLP & Social Media - Examples



(source)

HUMOR/IRONY/SARCASM



EMOTION

ABUSIVE LANGUAGE

SENTIMENT

ASPECT

Nice view of river. But sushi rolls are bad.

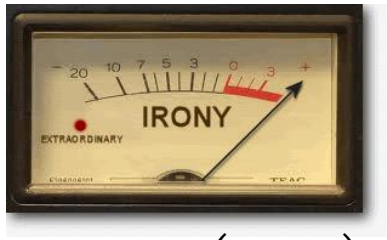
Sentiment expression | Aspect expression of LOCATION | Aspect expression of FOOD | Sentiment expression

(source)

STANCE

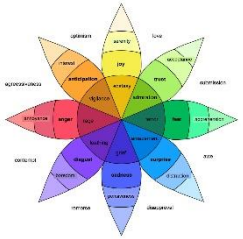
MIS/DISINFORMATION

NLP & Social Media - Examples



(source)

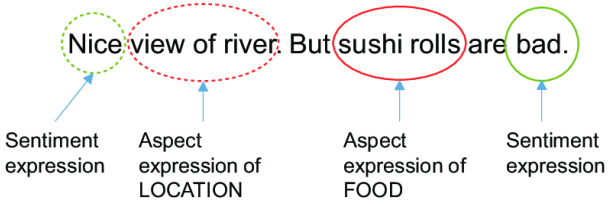
HUMOR/IRONY/SARCASM



EMOTION

SENTIMENT

ASPECT



(source)

STANCE

ABUSIVE LANGUAGE

- PROFANITITES
- STEREOTYPES
- OFFENSIVE L.
- AGGRESSION
- HATE SPEECH

MIS/DISINFORMATION

- FAKE NEWS
- PROPAGANDA TECHNIQUES

Resources for Italian

EVALITA

Evaluation of NLP and Speech Tools for Italian

EVALITA

Evaluation of NLP and Speech Tools for Italian

ABUSIVE LANGUAGE

HUMOR/IRONY/SARCASM

ASPECT

SENTIMENT → sentipolc@evalita 2014 + sentipolc@evalita 2016

STANCE

sentipolc@evalita 2014
+
sentipolc@evalita 2016
+
ironita@evalita 2018

HUMOR/IRONY/SARCASM

EVALITA
Evaluation of NLP and Speech Tools for Italian

ABUSIVE LANGUAGE

ASPECT

SENTIMENT → sentipolc@evalita 2014
+
sentipolc@evalita 2016

STANCE

sentipolc@evalita 2014
+
sentipolc@evalita 2016
+
ironita@evalita 2018

HUMOR/IRONY/SARCASM

EVALITA
Evaluation of NLP and Speech Tools for Italian

ABUSIVE LANGUAGE

ASPECT

ABSITA

(2018)

SENTIMENT

STANCE

sentipolc@evalita 2014
+
sentipolc@evalita 2016

sentipolc@evalita 2014
+
sentipolc@evalita 2016
+
ironita@evalita 2018

HUMOR/IRONY/SARCASM

EVALITA
Evaluation of NLP and Speech Tools for Italian

ABUSIVE LANGUAGE

ASPECT

ABSITA

(2018)

SENTIMENT

sentipolc@evalita 2014
+
sentipolc@evalita 2016

STANCE

SardiStance

(2020)

EVALITA

Evaluation of NLP and Speech Tools for Italian

haspeede2@evalita 2020

STEREOTYPES

ABUSIVE LANGUAGE

HUMOR/IRONY/SARCASM

HATE SPEECH

ASPECT

SENTIMENT

STANCE

AMI

Automatic Misogyny Identification – EVALITA 2018

+

Automatic Misogyny Identification (AMI)

Shared Task at EVALITA 2020

haspeede@evalita 2018

+

haspeede2@evalita 2020

Universal Dependencies

Italian	8	858K	W	IE, Romance

Italian treebanks

VIT	279K	LF		★★★★☆
ParTUT	55K	LF	W	★★★½☆
Valico	6K	LF		★★★½☆
MarkIT	40K	LF		★★½☆☆
PUD	23K	LF	W	★★☆☆☆
ISDT	298K	LFD	W	☆☆☆☆☆
TWITTIRO	29K	LF		☆☆☆☆☆
PoSTWITA	124K	LF		☆☆☆☆☆

Universal Dependencies

Italian	8	858K	W		IE, Romance

Italian treebanks

VIT	279K	LF			★★★★☆
ParTUT	55K	LF	W		★★★½☆
Valico	6K	LF			★★★½☆
MarkIT	40K	LF			★★½☆☆
PUD	23K	LF	W		★★☆☆☆
ISDT	298K	LFD	W		☆☆☆☆☆
TWITTIRO	29K	LF			☆☆☆☆☆
PoSTWITA	124K	LF			☆☆☆☆☆

Irony + syntax

Universal Dependencies

Italian	8	858K	W		IE, Romance

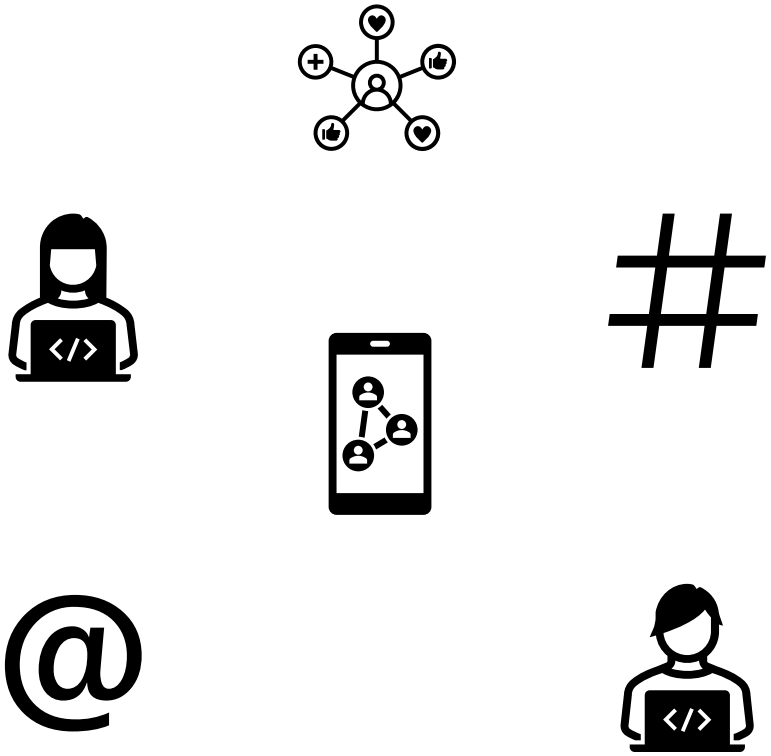
Italian treebanks

VIT	279K	LF			★★★★☆
ParTUT	55K	LF	W		★★★½☆
Valico	6K	LF			★★★½☆
MarkIT	40K	LF			★★½☆☆
PUD	23K	LF	W		★★☆☆☆
ISDT	298K	LFD	W		☆☆☆☆☆
TWITTIRO	29K	LF			☆☆☆☆☆
PoSTWITA	124K	LF			☆☆☆☆☆

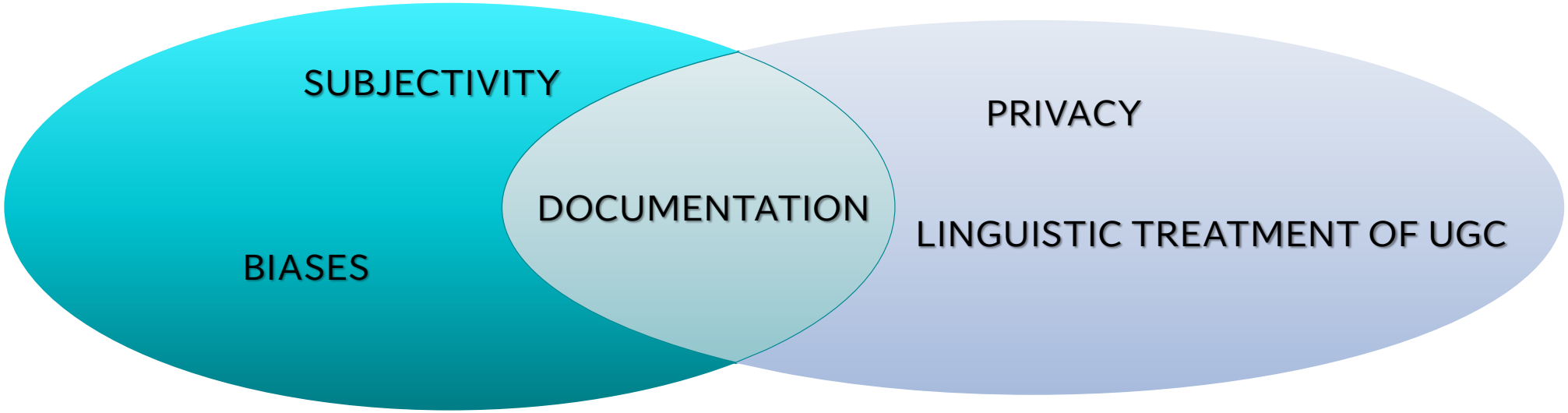
Partial overlap w/ Twittiro

Overview

1. NLP & Social Media
2. **Challenges & Issues in Resource Development**
3. Guidelines for UGC Data
4. Hands-on Session



Challenges & Issues in Resource Development



Subjectivity

- The usual annotation methodologies work well for traditionally relevant tasks in NLP, that typically rely on GOLD STANDARDS (i.e., *"the* **truth** *against which compare future predictions on the same set of instances"*)

- Critical issues are surfacing when applying old techniques to the study of highly subjective phenomena such as irony and sarcasm, or abusive and offensive language

(from Basile, 2021)

STEREOTYPE

?

Senza biglietto: nigeriana prende a testate e pugni capotreno

SARCASM

?

“I profughi che non arrivano più da un anno e mezzo” La gente deve essere proprio cretina per votare Salvini allora!

HATE SPEECH

?

Milano, si difendono da immigrati: comune li multa a 15 mila euro.. guardate il – VIDEO Commenta la #Mussolini VERGOGNAAAAA

Issues and Recommendations

- Low Inter-Annotator Agreement (usually seen as an upper bound of computer performance on the same task)
- Failure in recognizing «gold» labels as real-world objects for which there may not be a single truth
- Aggregation and harmonization destroy any personal opinion that comes as a result of the different cultural and demographic background of the annotators

ONGOING DISCUSSION (see the Perspectivist Data Manifesto):

- Create and distribute **non-aggregated** datasets
- Avoid evaluating models against aggregated gold standards

Biases

By “predictive bias," we refer to a situation in which a [predictive model] is used to predict a specific criterion for a particular population, and is found to give **systematically different predictions for subgroups of this population** who are in fact identical on that specific criterion

(from Shah et al, 2020)

bias noun

Save Word

bi·as | \ ˈbī-əs \

Definition of *bias* (Entry 1 of 4)

1 a : an inclination of temperament or outlook
especially : a personal and sometimes unreasoned judgment : PREJUDICE

b : an instance of such prejudice

c : BENT, TENDENCY

Challenges & Issues in Resource Development - Biases

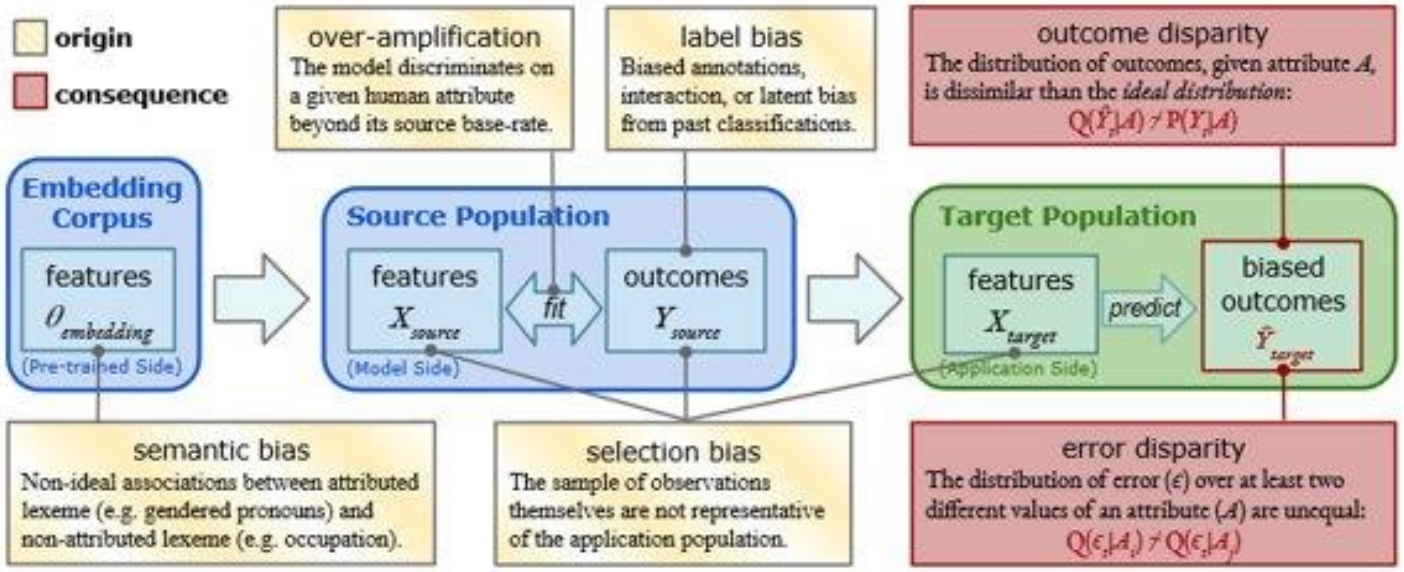


Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

(from Shah et al, 2020)

Challenges & Issues in Resource Development - Biases

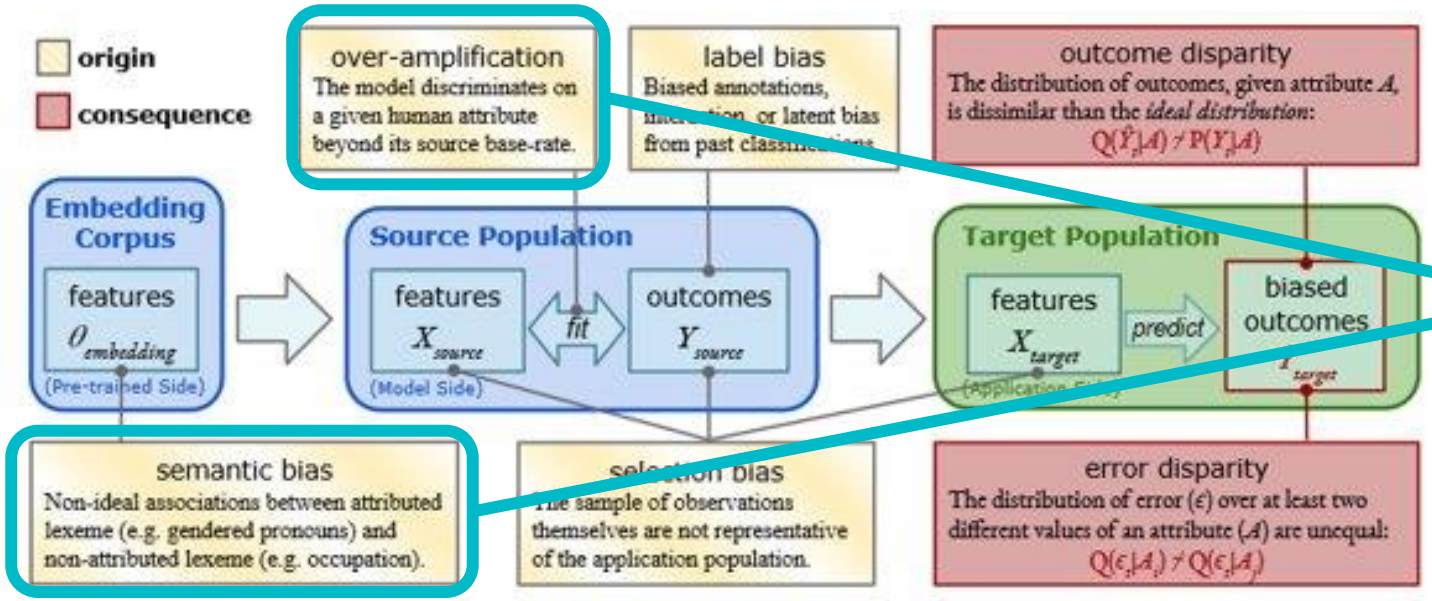
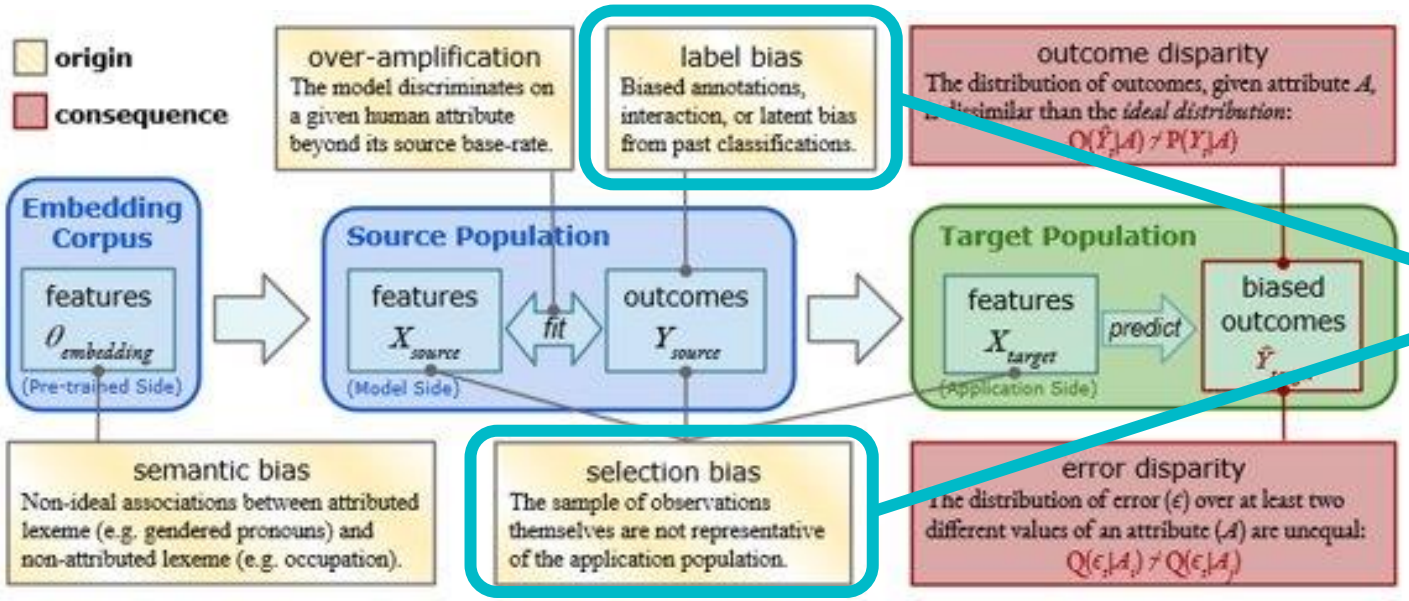


Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

Biased model

Challenges & Issues in Resource Development - Biases



Biased data

Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

Challenges & Issues in Resource Development - Biases

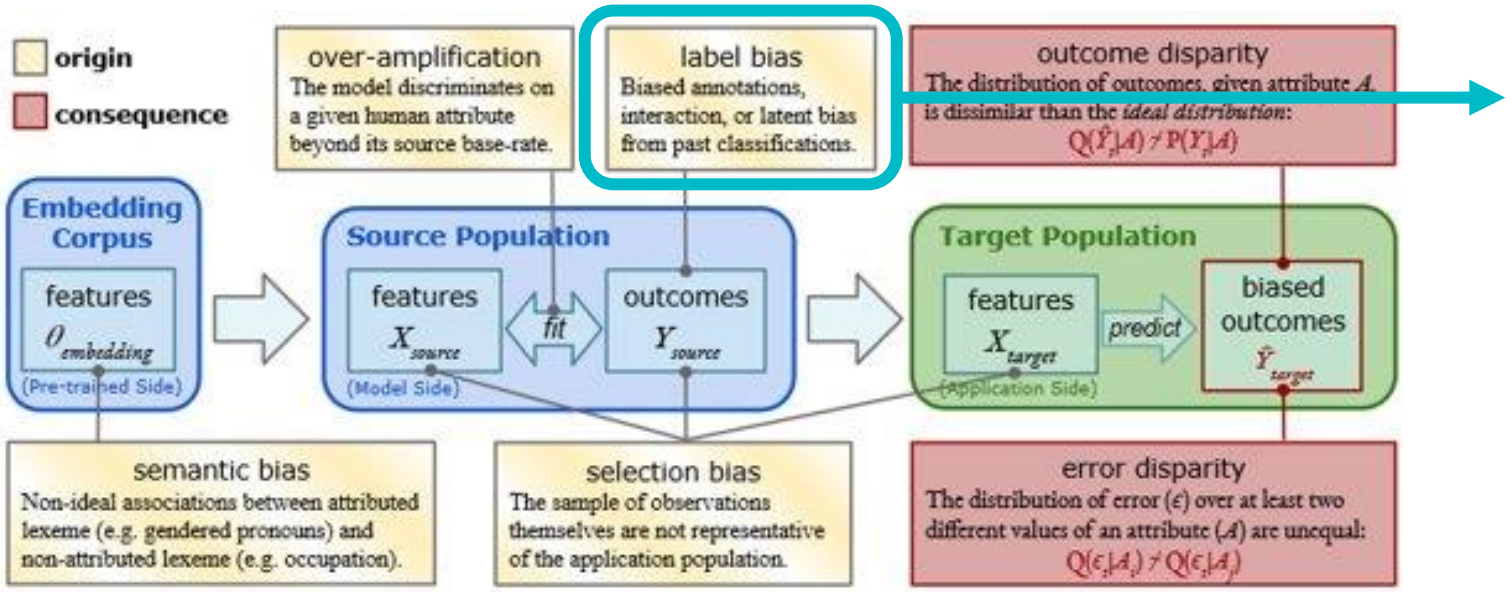


Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

WHY?

- non-representative group of annotators
- lack of domain expertise
- preconceived notions and stereotypes held by the annotators

Challenges & Issues in Resource Development - Biases

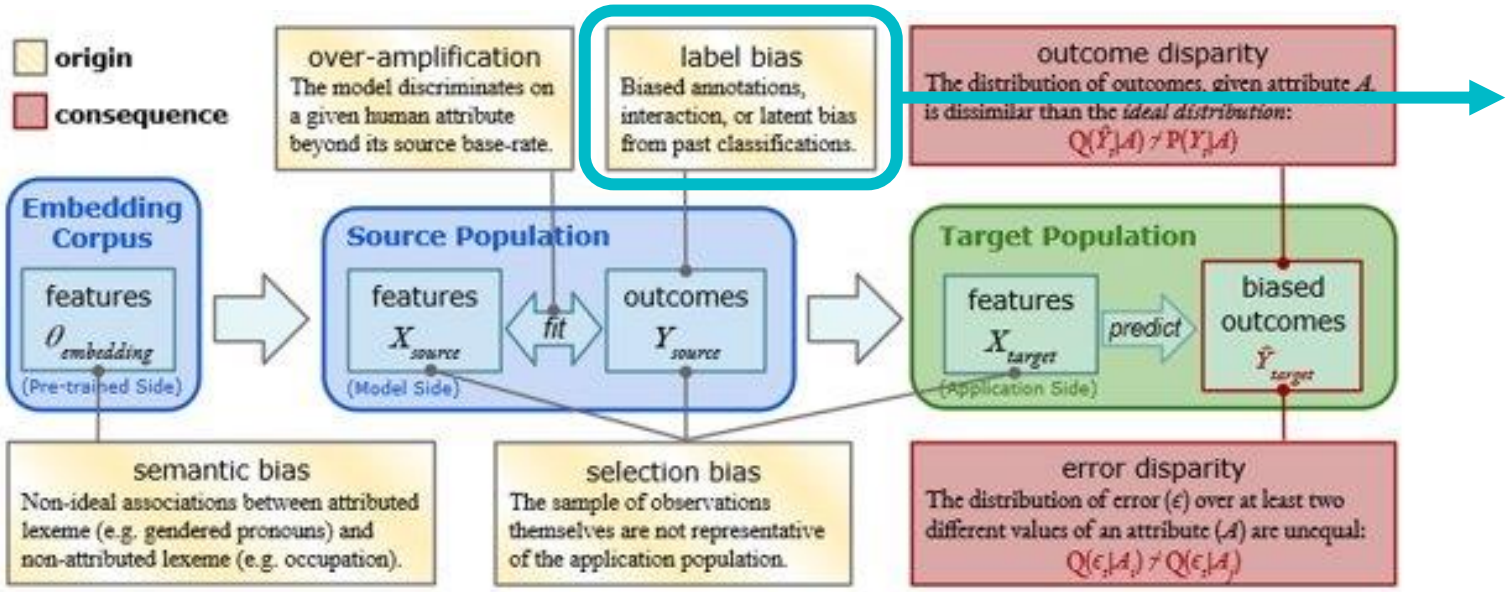


Figure 1: *The Predictive Bias Framework for NLP*: Depiction of where bias may originate within a standard supervised NLP pipeline. Evidence of bias is seen in $\hat{y}$ via *outcome disparity* and *error disparity*.

EXAMPLE:

- Disproportionate association between words describing queer identities and text labeled as “toxic” in a dataset for toxicity classification

Implications

ENVISAGED SCENARIO: Online fora may use NLP model for abusive language detection to censor undesirable language and promote civil discourse. Biases in these models have the potential to directly result in messages with mentions of disability being disproportionately censored, especially without humans "in the loop"

RESULT:

- Negative impact on people with disabilities and their opportunity to participate equally in online fora
- Readers and searchers of online fora might see fewer mentions of disability, exacerbating the already reduced visibility of disability in the public discourse

(from Hutchinson et al. 2020)

Recommendations

- Adopt possible countermeasures to mitigate biases (in both model and data). Among these:
  - Model induction from annotated data that take inter-annotator agreement into consideration (~~**label bias**~~)
  - Thorough documentation of the whole pipeline of dataset creation (see next slides)

Documentation

Increased need to get well-documented datasets (specially, bot not only, to minimize risks from biased data)

BUT

- Not often the case
- No standard guidelines

Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

Two well-known proposals

Recommendations

From Bender & Friedman, 2018:

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

- Curation rationale
- Language variety
- Speaker demographic
- Annotator(s) demographic
- Text characteristics
- Recording quality (for audio/video)
- Other
- Provenance appendix (for datasets built out of existing datasets)

Recommendations

From Bender & Friedman, 2018:

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

- Curation rationale PRIVACY? SELECTION BIAS?
- Language variety SELECTION BIAS?
- Speaker demographic SELECTION BIAS?
- Annotator(s) demographic LABEL BIAS?
- Text characteristics SELECTION BIAS?
- Recording quality (for audio/video)
- Other
- Provenance appendix (for datasets built out of existing datasets)

Privacy

- Due to the implications that the release of the data may have on the privacy of people, rules for their protection must be laid down
- These rules have been defined:
  - By the GDPR (within the EU context)
  - By the Terms of Service of the Web platforms

They provide legal bases for the collection-distribution-use of the data

Recommendations

Principle of data minimization (no more data than necessary) implemented through:

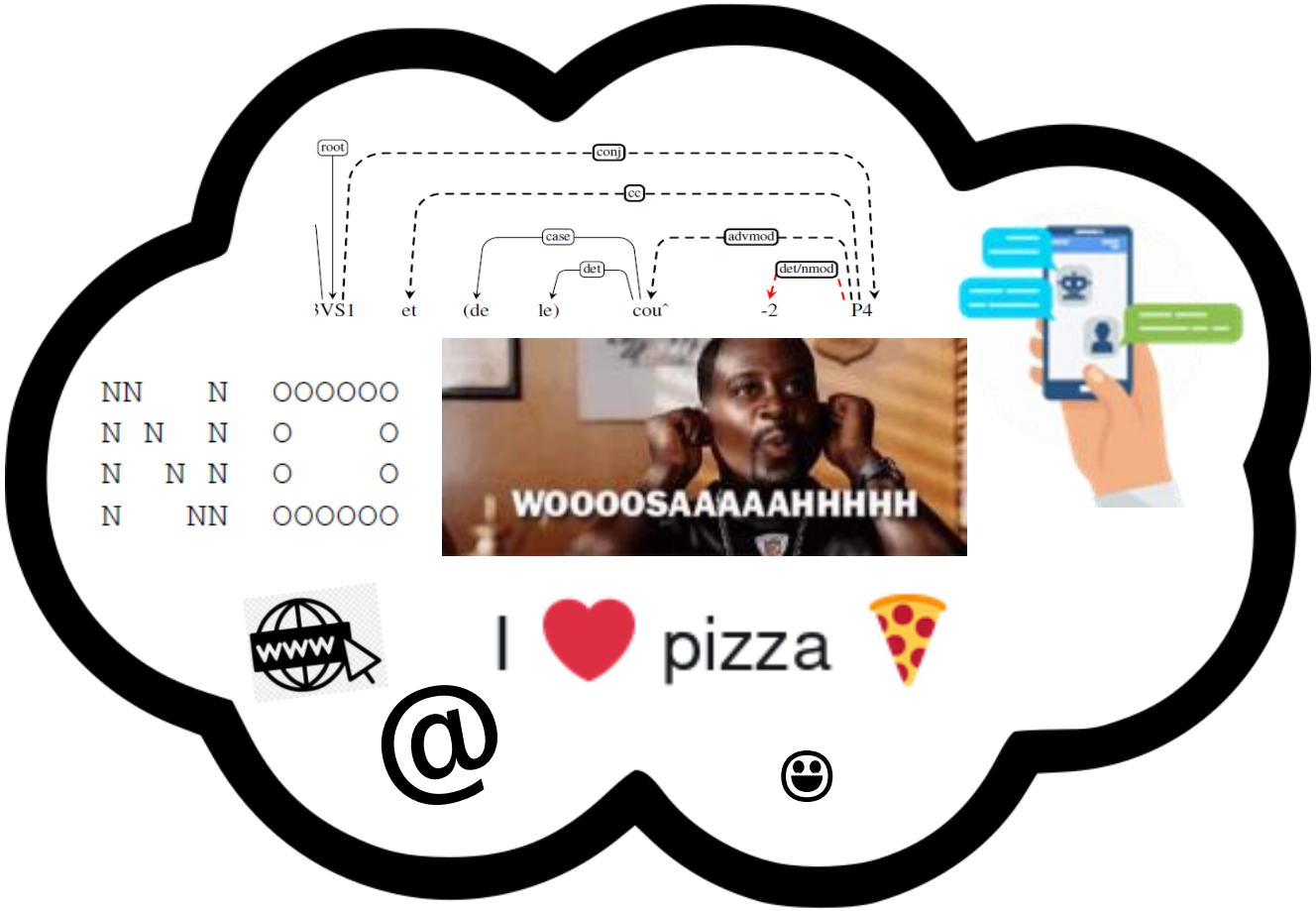
- Pseudonymization ⟶ hide personal data (e.g., user handle)
- Encryption ⟶ use password/control data access

Trade-off between legal compliance and research activity (providing access to data and maintain the reproducibility of the experiments)

(from Rangel and Rosso, 2019)

Linguistic Treatment of UGC Data

UGC is **noisy**!

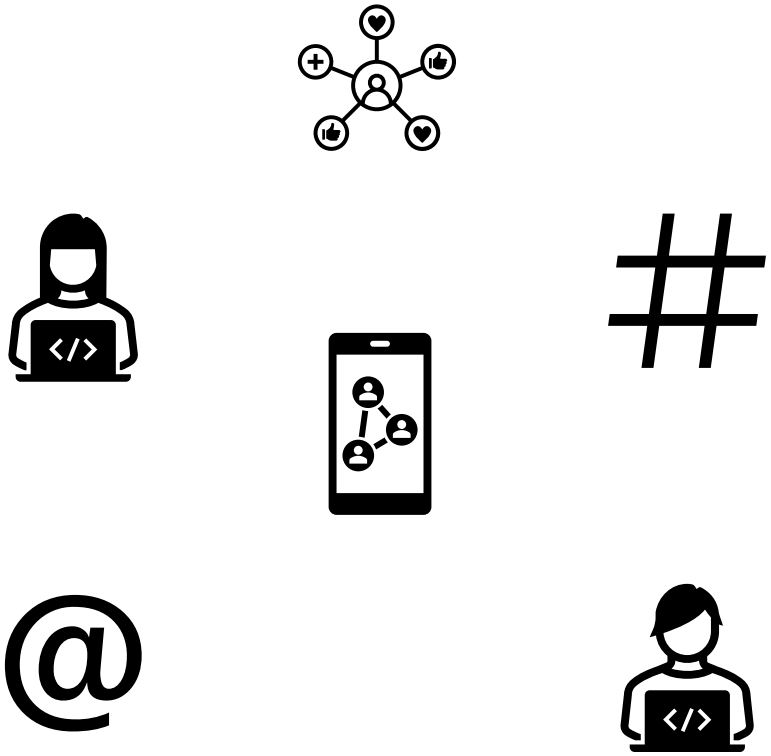


Recommendations

See next part of the talk...

Overview

1. NLP & Social Media
2. Challenges & Issues in Resource Development
3. **Guidelines for UGC Data**
4. Hands-on Session

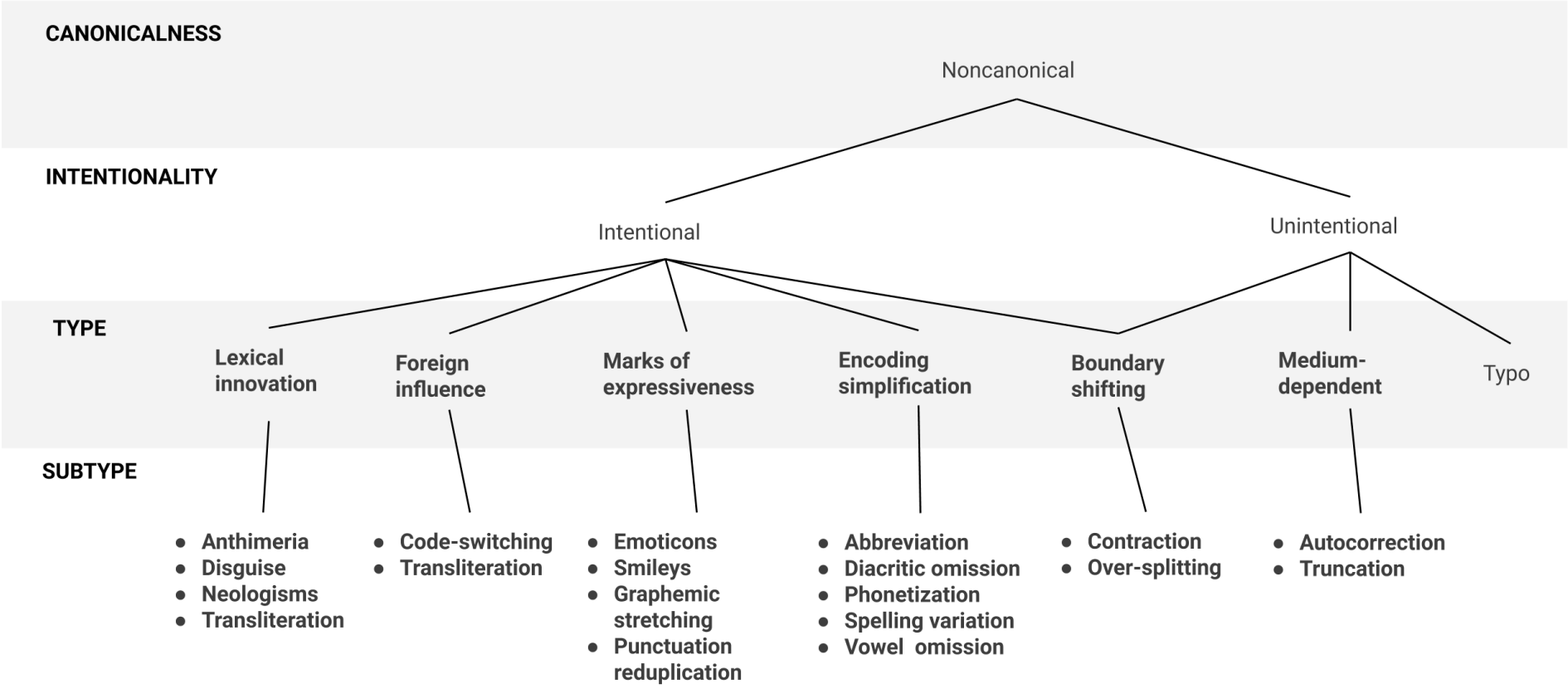


Guidelines for UGC Data

UGC is **noisy**!

- Spelling mistakes & typos
- Colloquialisms/slang/internet jargon
- Abbreviations & spelling variations
- Pictograms
- Non-standard syntactic constructions & ungrammatical text
- Code-mixing
- Embedded metadata (hashtags, URLs, mentions, ...)
- ...

A (tentative) Taxonomy



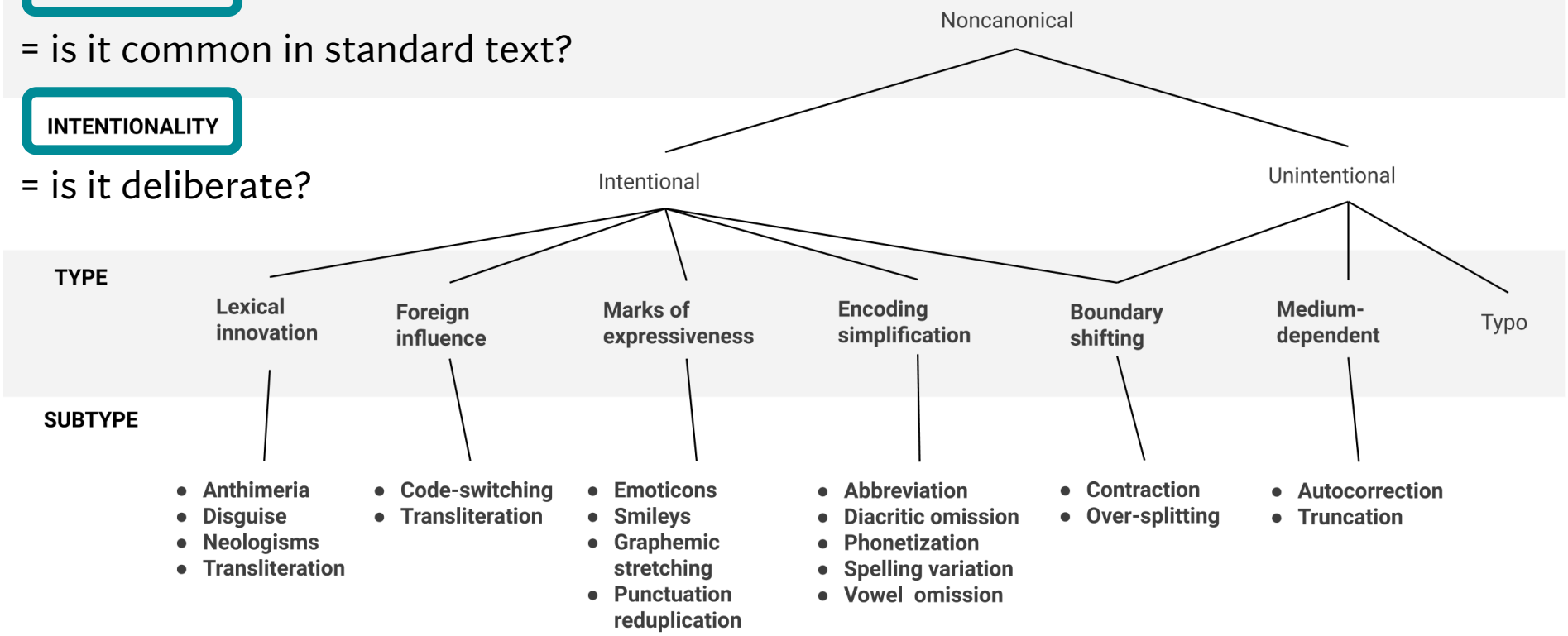
Guidelines for UGC Data – A (tentative) Taxonomy

CANONICALNESS

= is it common in standard text?

INTENTIONALITY

= is it deliberate?

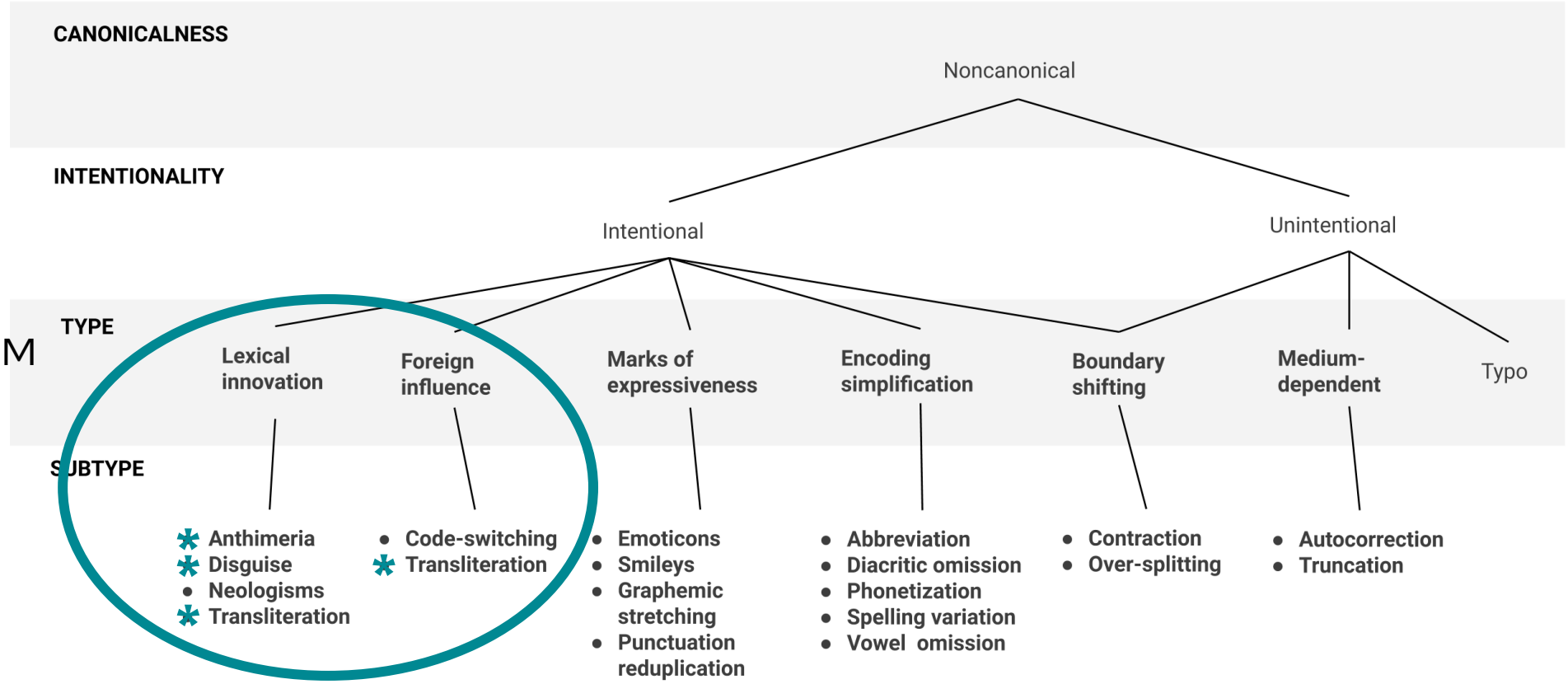


Guidelines for UGC Data – A (tentative) Taxonomy

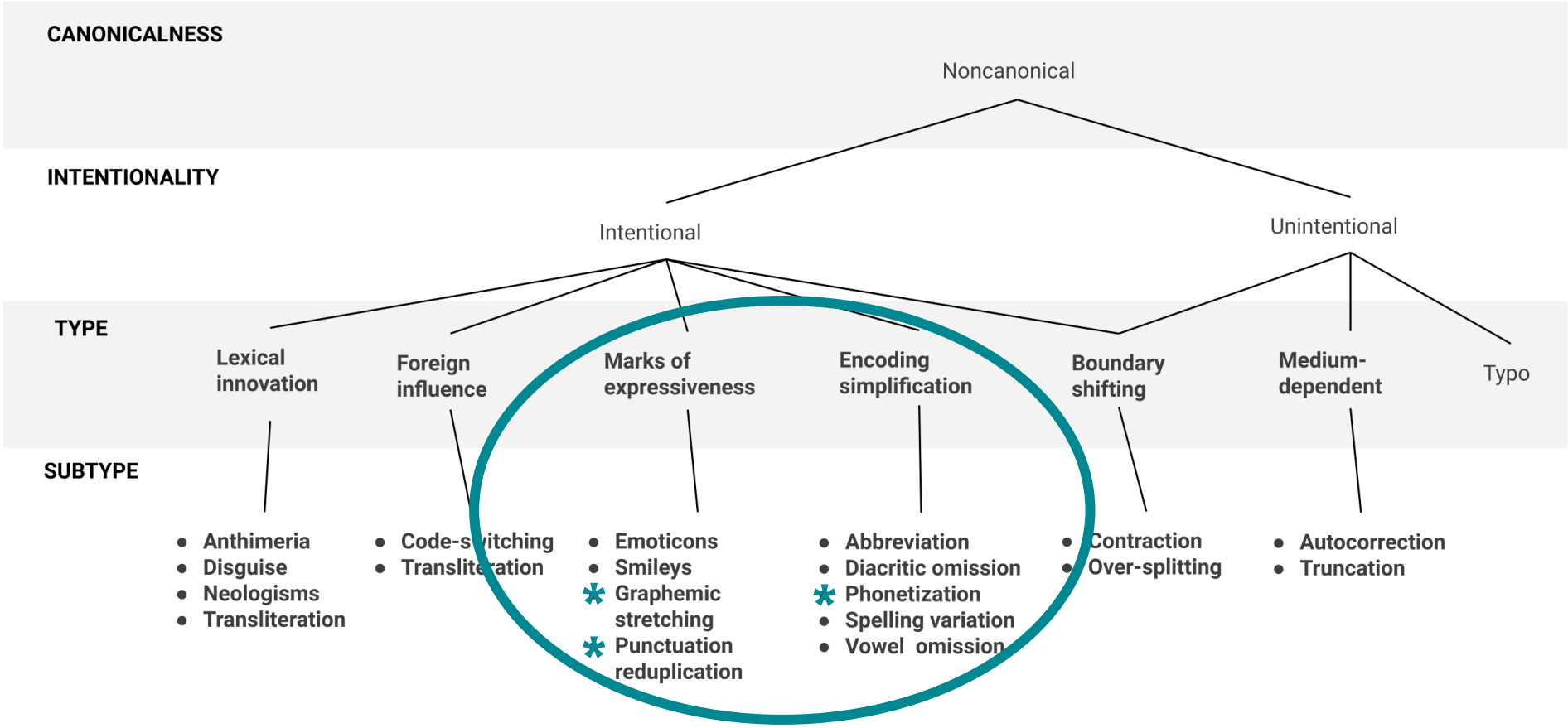
feel free to DM

caxxo!
(*f**k!*)

Féar plé
(fair play)



Guidelines for UGC Data – A (tentative) Taxonomy



chiiii?!??! (*whooo?!??!*)

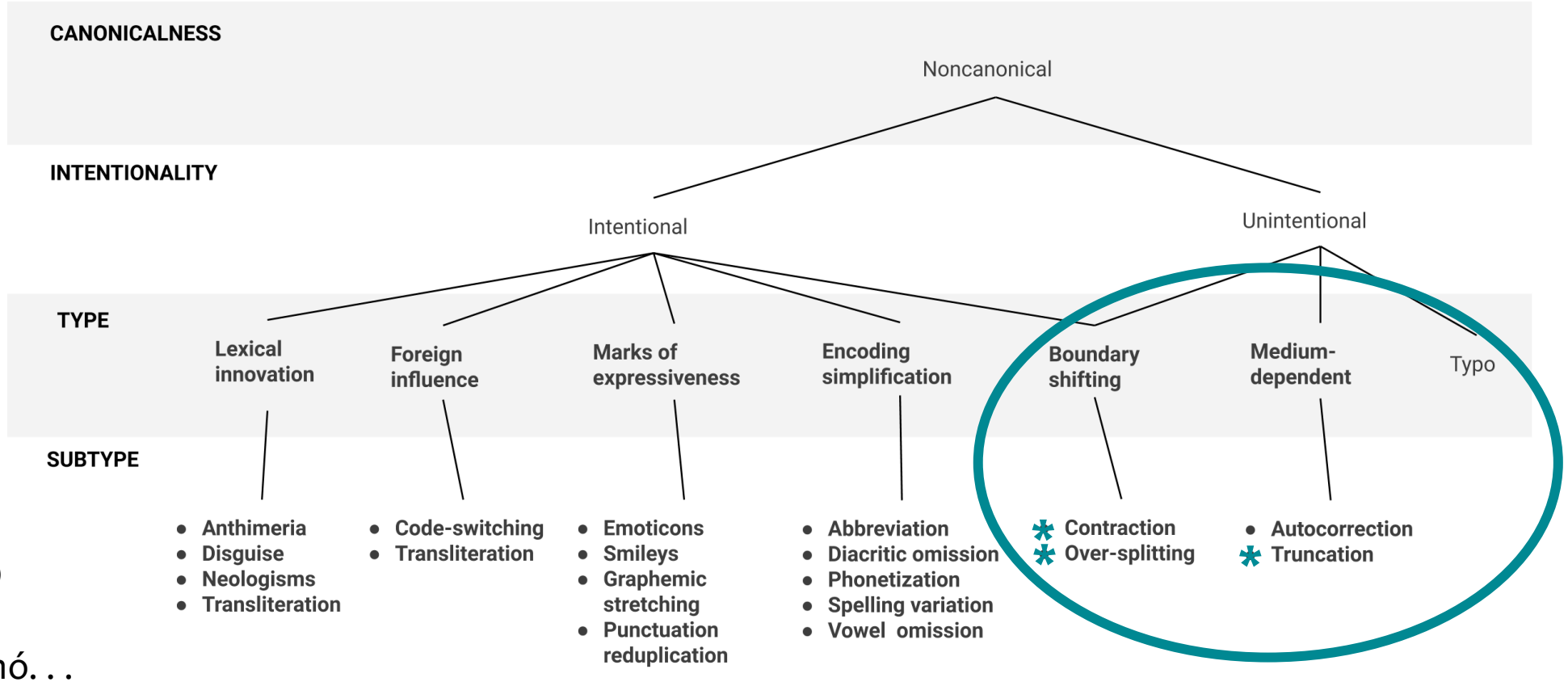
k1 Mensch (*no 1*)

Guidelines for UGC Data – A (tentative) Taxonomy

🇫🇷 nimp quoi
(*rubbish*)

🇹🇷 gele bilirim
(gelebilirim,
I-can-come)

🇮🇪 thart fa' 53 nó. . .
(*over 53 mi...(minutes)*)



UGC Treebanks

- A large number of UGC treebanks developed in the last decade

Name	Reference	Source	Language
FSMB	Seddah et al. (2012)	Twitter, Facebook discussions fora	FR
Narabizi (NBZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR
EWT	Silveira et al. (2014)	various	EN
LAS-DisFo (LDF)	Taulé et al. (2015)	discussion fora	ES
MoNoise (MNo)	Van Der Goot and van Noord (2018)	Twitter	EN
STB	Wang et al. (2017)	discussion fora	SgE
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH
GUM	Zeldes (2017)	various	EN
HSE	n.a.	various	BE
OOD	n.a.	various	FI
TwittIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA
Cadhan (Cdh)	n.a.	various	GV
Taiga	n.a.	various	RU
IU	n.a.	various	UK

Name	Reference	Source	Language
ATDT	Albogamy and Ramsay (2017)	Twitter	AR
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE
TWITTIRÒ-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT
DWT	Daiber and Van Der Goot (2016)	Twitter	EN
W2.0	Foster et al. (2011)	Twitter, sport fora	EN
Foreebank (Frb)	Kaljahi et al. (2015)	technical fora	EN, FR
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN
TDT	Luotolahti et al. (2015)	various	FI
xUGC	Martínez Alonso et al. (2016)	various	FR
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	various	ET
ITU	Pamay et al. (2015)	n.a.	TR
WDC	Read et al. (2012b)	various	EN
tweeDe	Rehbein et al. (2019)	Twitter	DE
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT

- Most of them are UD-based, but still adopt different strategies

Name	Reference	Source	Language
FSMB	Seddah et al. (2012)	Twitter, Facebook discussions fora	FR
Narabizi (NBZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR
EWT	Silveira et al. (2014)	various	EN
LAS-DisFo (LDF)	Taulé et al. (2015)	discussion fora	ES
MoNoise (MNo)	Van Der Goot and van Noord (2018)	Twitter	EN
STB	Wang et al. (2017)	discussion fora	SgE
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH
GUM	Zeldes (2017)	various	EN
HSE	n.a.	various	BE
OOD	n.a.	various	FI
TwittIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA
Cadhan (Cdh)	n.a.	various	GV
Taiga	n.a.	various	RU
IU	n.a.	various	UK

Name	Reference	Source	Language
ATDT	Albogamy and Ramsay (2017)	Twitter	AR
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE
TWITTIRÒ-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT
DWT	Daiber and Van Der Goot (2016)	Twitter	EN
W2.0	Foster et al. (2011)	Twitter, sport fora	EN
Foreebank (Frb)	Kaljahi et al. (2015)	technical fora	EN, FR
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN
TDT	Luotolahti et al. (2015)	various	FI
xUGC	Martínez Alonso et al. (2016)	various	FR
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	various	ET
ITU	Pamay et al. (2015)	n.a.	TR
WDC	Read et al. (2012b)	various	EN
tweeDe	Rehbein et al. (2019)	Twitter	DE
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT

Towards a Unified Representation

- A UD-based annotation framework that recommends stategies and guidelines for each linguistic level taken into account in the Universal Dependencies:

LEVEL	STRATEGY
TOKEN	MERGED *vs* SPLIT
LEMMA	NORMALIZED *vs* AS-IS
UPOS – FEATS – DEPREL	INTEGRATED *vs* STANDALONE

General Principles

- Strike a balance between **feasibility** and **consistency**

LEVEL	STRATEGY
TOKEN	MERGED *vs* SPLIT
LEMMA	NORMALIZED *vs* AS-IS
UPOS – FEATS – DEPREL	INTEGRATED *vs* STANDALONE

LEVEL	STRATEGY
TOKEN	MERGED *vs* SPLIT
LEMMA	NORMALIZED *vs* AS-IS
UPOS – FEATS – DEPREL	INTEGRATED *vs* STANDALONE

- Normalize/resort to more standardized forms whenever possible

LEVEL	STRATEGY
TOKEN	MERGED *vs* SPLIT
LEMMA	NORMALIZED *vs* AS-IS
UPOS – FEATS – DEPREL	INTEGRATED *vs* STANDALONE

- Establish whether the token bears a semantic/syntactic role or not

Summary

	Merged	Split	Normalized	As-is	Integrated	Standalone
Spelling mistakes & typos						
Punctuation/Spelling variation						
Contractions/Abbreviations						
Neologisms/slang						
Pictograms						
Metadata (#, @, URLs)						
Code-mixing						

	Merged	Split	Normalized	As-is	Integrated	Standalone
Spelling mistakes & typos						
Punctuation/Spelling variati						
Contractions/Abbreviations						
Neologisms/slang						
Pictograms						
Metadata (#, @, URLs)						
Code-mixing						

DISCLAIMER

The following summary is a rough approximation of what proposed in the UGC guidelines. It provides a generalization of the main guiding principles that motivated our proposal, and it leaves out many specific cases that would have required a more in-depth discussion.

Guidelines for UGC Data - Summary

	TOKEN					
	Merged	**Split**	**Normalized**	**As-is**	**Integrated**	**Standalone**
Spelling mistakes & typos	(UD guidelines)					
Punctuation/Spelling variation		✕				
Contractions/Abbreviations		✕				
Neologisms/slang	(context-dependent)					
Pictograms		✕				
Metadata (#, @, URLs)		✕				
Code-mixing	(doesn't apply)					

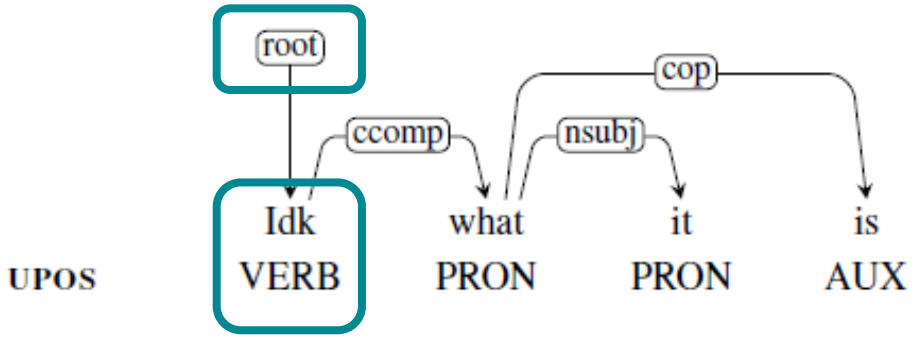
Guidelines for UGC Data - Summary

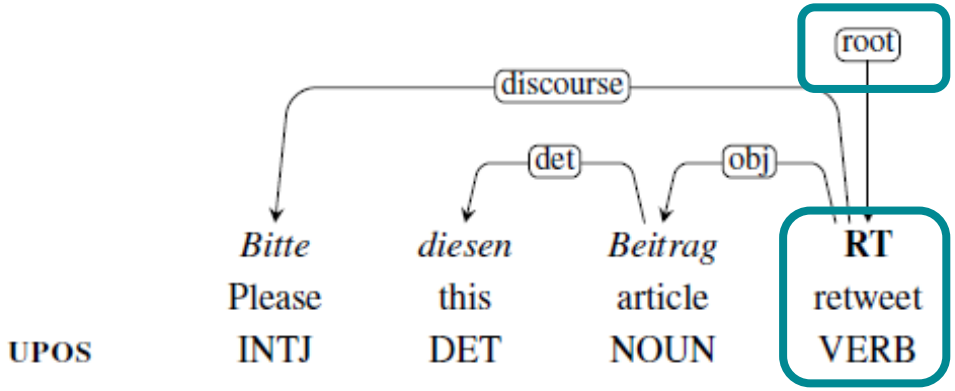
	TOKEN		LEMMA			
	Merged	**Split**	**Normalized**	**As-is**	**Integrated**	**Standalone**
Spelling mistakes & typos	(UD guidelines)		✓			
Punctuation/Spelling variation		✗	✓			
Contractions/Abbreviations		✗		✓		
Neologisms/slang	(context-dependent)			✓		
Pictograms		✗	(context-dependent)			
Metadata (#, @, URLs)		✗		✓		
Code-mixing	(doesn't apply)		(annotator-dependent)			

Guidelines for UGC Data - Summary

	TOKEN		LEMMA		UPOS – FEATS – DEPRELS	
	Merged	**Split**	**Normalized**	**As-is**	**Integrated**	**Standalone**
Spelling mistakes & typos	(UD guidelines)		✓		(UD guidelines)	
Punctuation/Spelling variation		✗	✓		(UD guidelines)	
Contractions/Abbreviations		✗		✓	✓	
Neologisms/slang	(context-dependent)			✓	(UD guidelines)	
Pictograms		✗	(context-dependent)		(context-dependent)	
Metadata (#, @, URLs)		✗		✓	(context-dependent)	
Code-mixing	(doesn't apply)		(annotator-dependent)		✓	

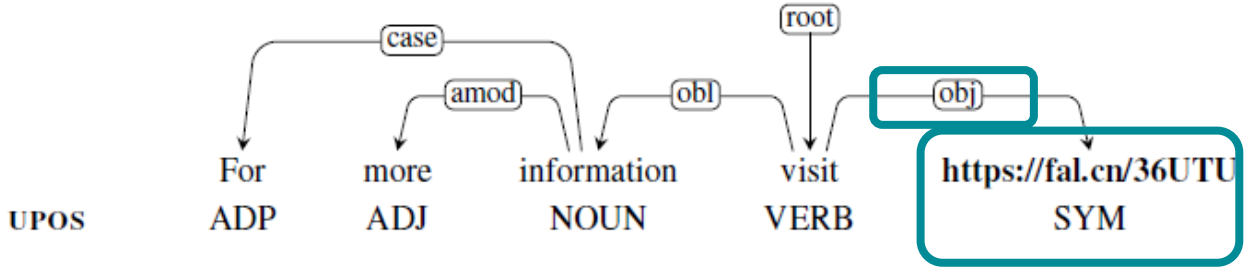
Contractions & Abbreviations

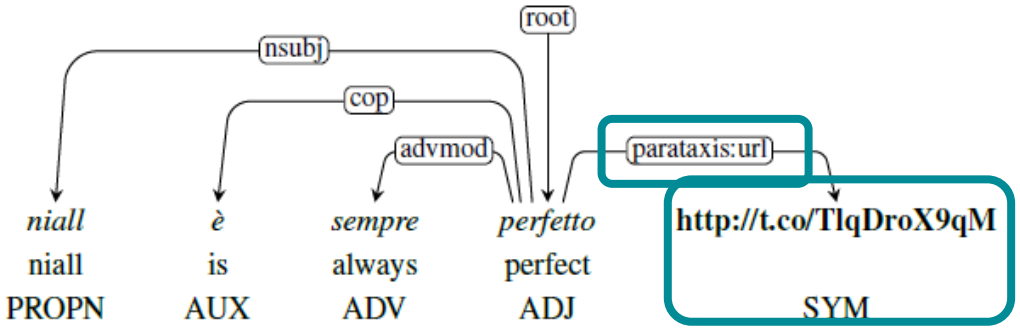




'Please retweet this article.'

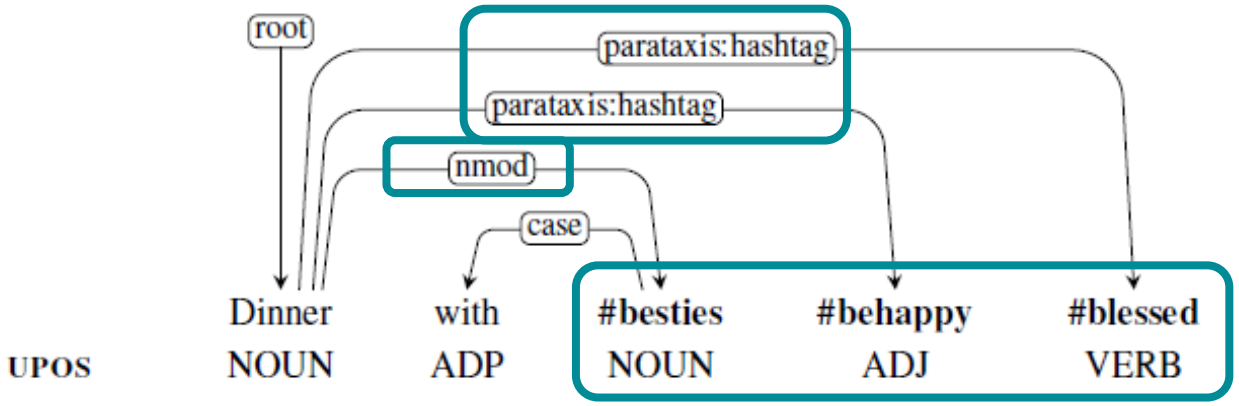
Embedded Metadata

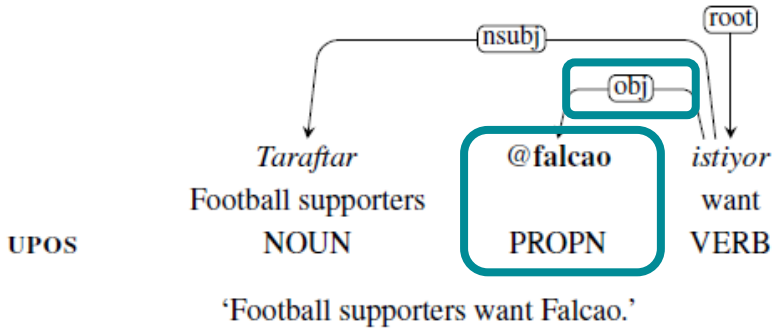


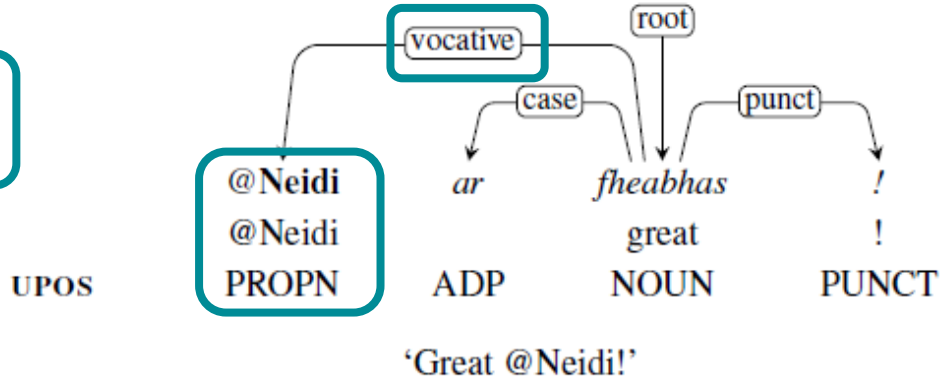


'Niall is always impeccable'

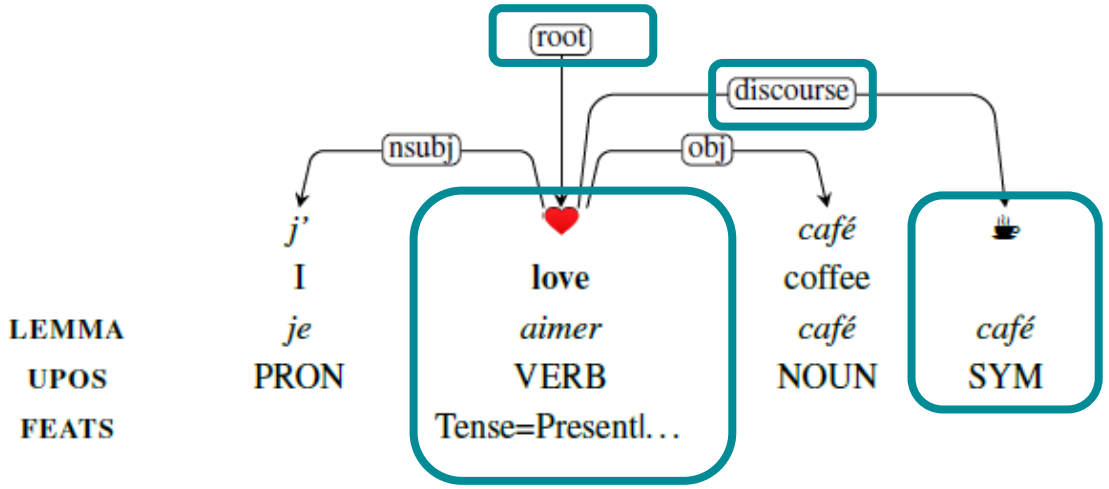
Embedded Metadata



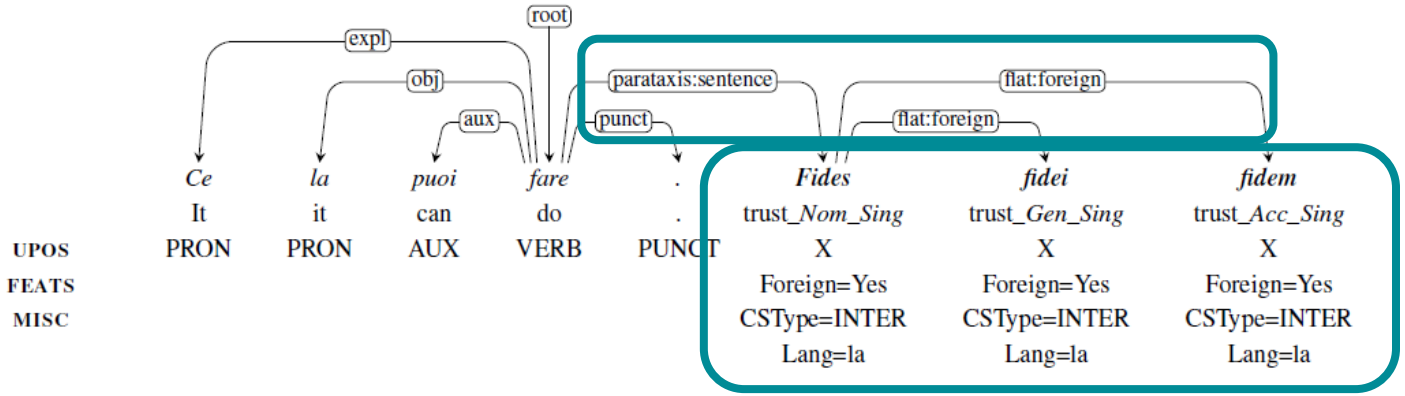




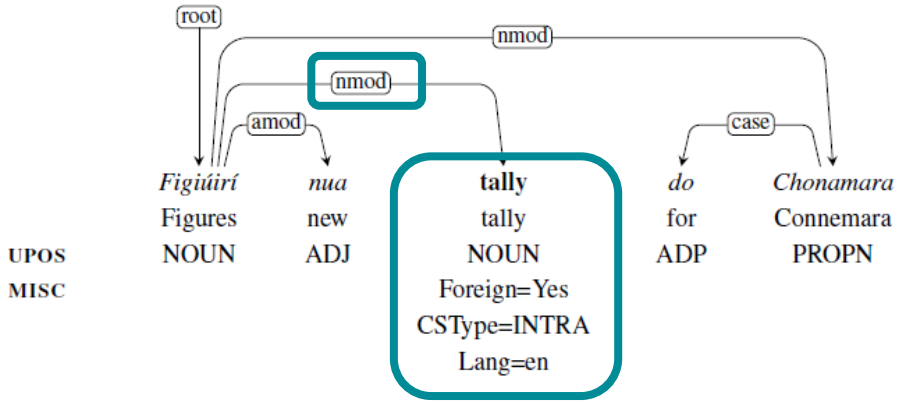
Pictograms


root
discourse
nsubj
obj
j'
I
je
PRON
love
aimer
VERB
Tense=Presentl...
café
coffee
café
NOUN
café
SYM
LEMMA
UPOS
FEATS

Code-mixing



'You can do it. Trust trust trust...'



'New tally figures for Connemara'

The complete guidelines, with other examples and summarizing tables,

can be found in the paper

Treebanking User-Generated Content: A UD-Based Overview of Guidelines, Corpora and Unified Recommendations

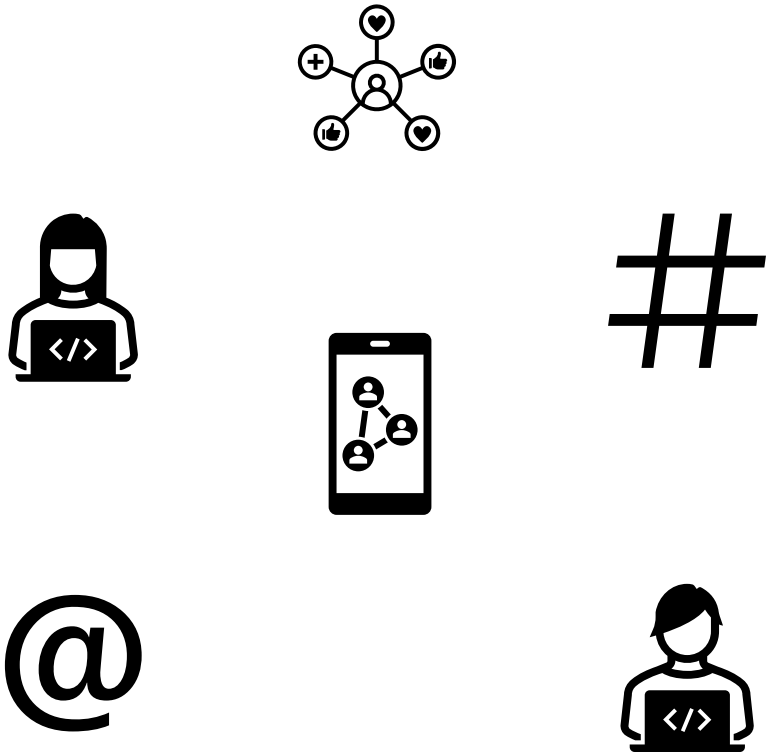
also available in this repository

https://github.com/msang/seminarUniTo/

Questions?

Overview

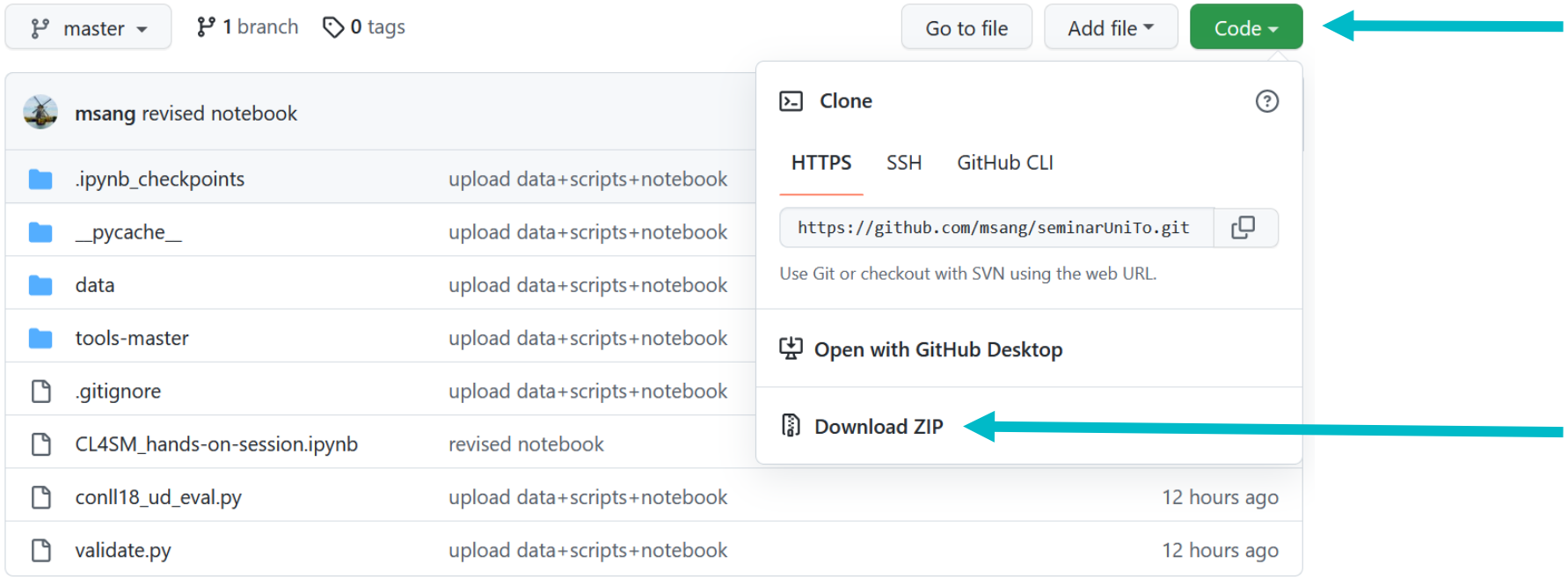
1. NLP & Social Media
2. Challenges & Issues in Resource Development
3. Guidelines for UGC Data
4. **Hands-on Session**



Hands-on Session

1. Download and unzip the folder from the GitHub repository:

https://github.com/msang/seminarUniTo



2. Upload the folder to Google Drive and run the notebook with Colab

OR

Run the project in your local machine:

```
conda install jupyter
```

or

```
pip install jupyter
```

Install Jupyter via command line

```
cd path_to_directory/
jupyter notebook
```

Move to the directory of the project and run the notebook

3. Follow the instructions in the notebook

Part I - Select & Parse

- Pick a text sample of your choice among the ones available in this repository
- Parse your data through the UDPipe web service

Part II - Revise

- Save a copy of the UDPipe output file (`processed.conllu`) as `revised.conllu`
- Manually revise the parsed data in the `revised.conllu` file using a text editor, or with a GUI (e.g., Inception)

  📋 When in doubt, feel free to consult the UD main guidelines, or the proposed guidelines for the treatment of UGC data (here and here the summarizing tables)

  💡 For the sake of simplicity, save **both conllu files in the same directory** as this notebook
- Validate the file (to make sure it doesn't contain any formatting error)

  ⚠️ Make sure that the language flag (`--lang`) of the script has the proper ISO code: English = `en`, French = `fr`, German = `de`, Italian = `it`,

Part III - Evaluate

- Evaluate the parser's performance on your data using standard metrics: