TAGSETS AND DATASETS: SOME EXPERIMENTS BASED ON PORTUGUESE LANGUAGE

Cláudia Freitas (PUC-Rio)
Luiza F. Trugo (PUC-Rio)
Fabricio Chalub (IBM Research)
Guilherme Paulino-Passos (IBM Research)
Alexandre Rademaker (IBM Research)

MOTIVATION AND PURPOSES

"*What are the linguistic distinctions most effective/suitable to NLP understanding?*"

- Investigate the impact of different (PoS) tagsets on the performance of NLP systems

- Studies comparing PoS-taggers are frequent; linguistic studies comparing the impact of different tagsets on the performance of NLP systems are scarce.

- A possible reason for the imbalance may lie in the belief that the distribution of words along certain categories are based on *objective* and *stable properties* associated with words, and it is up to machines or programmers to develop the best classification strategy.

Requirement for this kind of (linguistic) study: comparable materials

- Same corpus; different tagsets
- Macmorpho corpus
  - Marcmorpho tagset
  - Universal Dependencies tagset

TWO EXPLORATORY STUDIES

Evaluate the performance of a learning model when the (pos) tagset is modified

- Better linguistic classes could help ML

Evaluate the performance of a learning model when we modify the *amount* of training (and test) material

- "side-effect" study

PREPARING THE ENVIRONMENT: SAME CORPUS; DIFFERENT TAGSET

Macmorpho (pos) tagset

MacMorpho Corpus v1 (Aluisio et al.,2003)

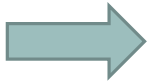
- 1.1 million words
- annotated at the PoS level
- manually revised
- widely used in Portuguese PoS training
- bigger than Bosque-UD (244,675 words)
  - Bosque-UD wasn't available at the time of the conversion process

MacMorpho v1 tagset: 23 labels

UD (pos) tagset

Universal Dependencies (UD)

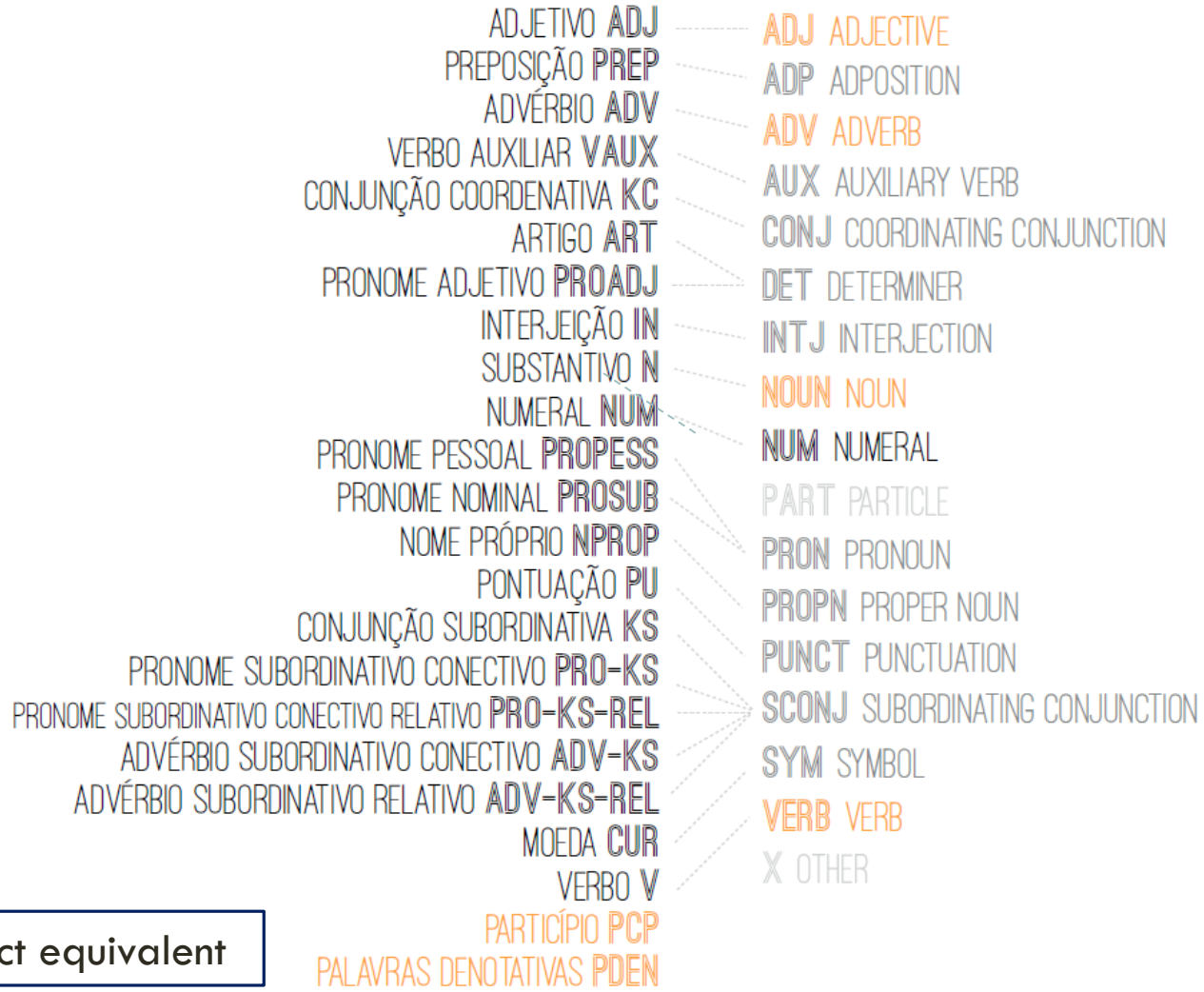
Framework for cross-linguistic grammatical annotation that aims at developing a language-independent annotation scheme, flexible for specific extensions of a given language tagset



UD 2.0 tagset: 17 labels

MACMORPHO (POS) TAGSET → UD (POS) TAGSET

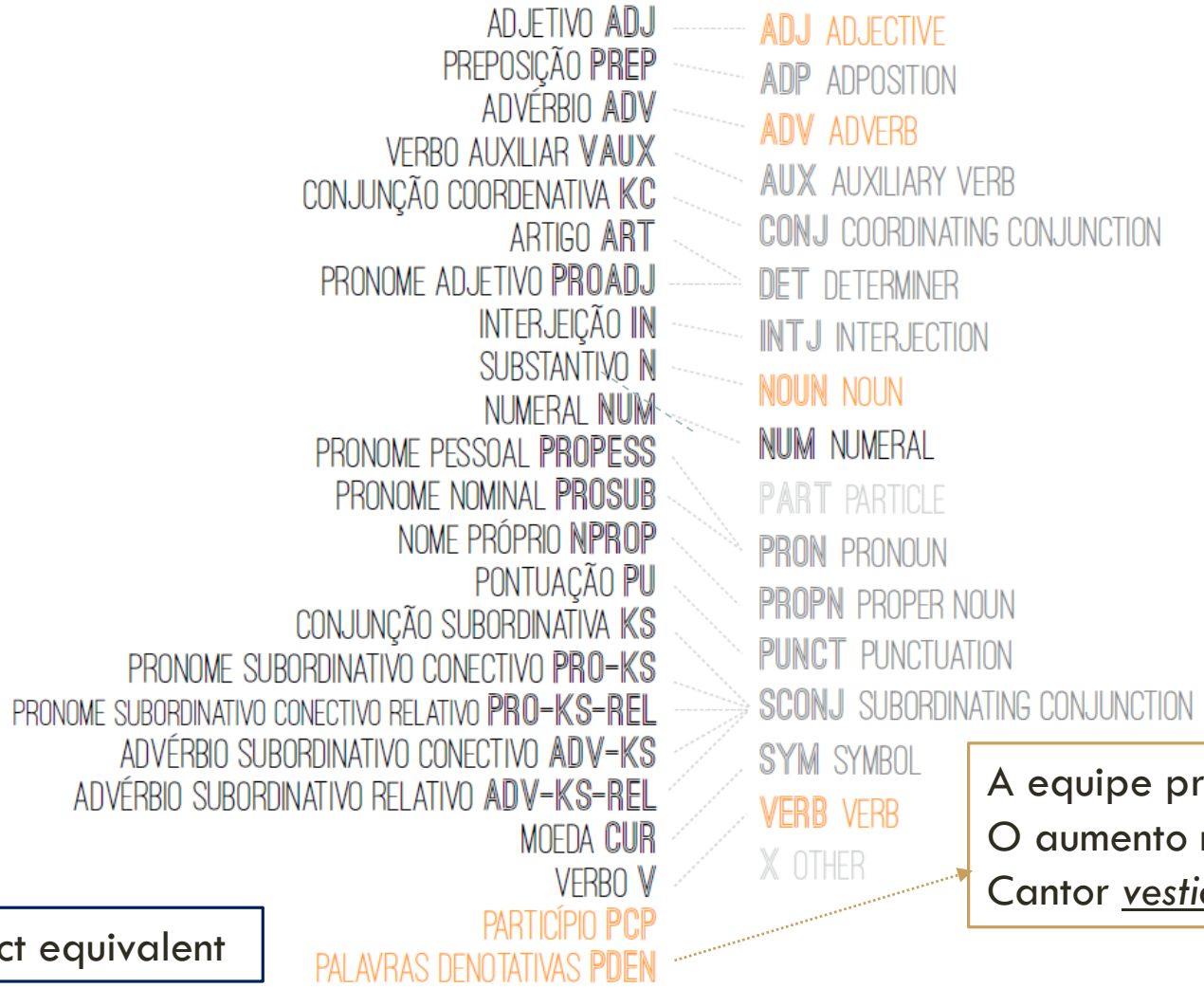
1. The same label can be used for different purposes:
Both MM and UD have NUM, for *Numbers*, but… in the MM corpus, if the *numeral* is functioning as the head of a NP, it should be tagged as NOUN.



2. There were tags without a direct equivalent

MACMORPHO (POS) TAGSET → UD (POS) TAGSET

1. The same label can be used for different purposes:
Both MM and UD have NUM, for *Numbers*, but… in the MM corpus, if the *numeral* is functioning as the head of a NP, it should be tagged as NOUN.



2. There were tags without a direct equivalent

A equipe precisa jogar *adiantado* (ADV)
O aumento no preço dos *importados* (NOUN)
Cantor *vestido* de Elvis (VERB? ADJ?)

CONVERSION – PCP... ADJ OR VERB?

"*Refiro-me mesmo àqueles programas de interesse mais geral, como as telenovelas, já que nem essas entram às horas anunciadas.*"

- ... às horas *que foram anunciadas* (VERB) ?
- ... às horas *certas*? (ADJ) ?

(*I'm referring to those programs of more general interest, such as soap operas, since neither do they start at the announced/expected hours.*)

- Uncertainty about the "right" tag
  - High level of disagreement (50% in some cases) (Trugo, 2016)

CONVERSION – PCP... ADJ OR VERB?

"*Refiro-me mesmo àqueles programas de interesse mais geral, como as telenovelas, já que nem essas entram às horas anunciadas.*"

... às horas *que foram anunciadas* (VERB) ?

... às horas *certas*? (ADJ) ?

(*I'm referring to those programs of more general interest, such as soap operas, since neither do they start at the announced/expected hours.*)

- Uncertainty about the "right" tag
  - High level of disagreement (50% in some cases) (Trugo, 2016)

Participle

When parts of speech were translated from Greek into Latin, the **participle** (***participium***) was named precisely for "participating" in two classes at the same time: NOMINALs and VERBs

- Participles were an independent class

- The same *phenomenon* appears in other languages:

"*One recurring area of difficulty, in all the languages for which we have been involved in lexicography – two recent examples being Polish and Estonian – is* ***participles /gerunds.***" (Kilgarriff & Kosem, 2012)

CONVERSION – TECHNICAL ASPECTS

A library in Common Lisp was developed in order to apply the rules (Muniz et al., 2017)

- Rules are stored independently of the corpus
- It produces not only the output data but also some detailed report of the rules applied to each sentence (log files).

CONVERSION - RESULTS

3 corpora:

- MM-UD (MacMorpho corpus + UD tagset)
- MM-UD+ PCP (MacMorpho corpus + UD tagset + PCPtag for *participle* forms)
- MM (*classic* MacMorpho corpus + some *light* revision)
  - ~~Deputado_PCP~~ → Deputado_N

TEST 1- WHAT HAPPENS WHEN WE USE DIFFERENT TAGSETS?

Dataset	MaxEnt Accuracy *
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

* Maximum Entropy model provided by the OpenNLP suite. Train and dev partitions were merged.

TEST 1- WHAT HAPPENS WHEN WE USE DIFFERENT TAGSETS?

Is the PCP responsible for these results?

Dataset	MaxEnt Accuracy *
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

* Maximum Entropy model provided by the OpenNLP suite. Train and dev partitions were merged.

TEST 1- WHAT HAPPENS WHEN WE USE DIFFERENT TAGSETS?

Is the PCP responsible for these results?

Dataset	MaxEnt Accuracy *
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

Comparison of confusion between the scenarios with and without PCP.

Golden PoS	Predicted PoS	Confusion	
		with PCP	without PCP
VERB	NOUN	188	261
NOUN	VERB	182	271
ADJ	VERB	**53**	**295**
VERB	ADJ	**32**	**287**

TEST 1- WHAT HAPPENS WHEN WE USE DIFFERENT TAGSETS?

Is the PCP responsible for these results?

Dataset	MaxEnt Accuracy *
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

The best performance with UD+PCP tagset is the result of the PCP label, wich plays the role of a disambiguator, artificially constructing a consensus

Comparison of confusion between the scenarios with and without PCP.

Golden PoS	Predicted PoS	Confusion	
		with PCP	without PCP
VERB	NOUN	188	261
NOUN	VERB	182	271
ADJ	VERB	**53**	**295**
VERB	ADJ	**32**	**287**

TEST 1- WHAT HAPPENS WHEN WE USE DIFFERENT TAGSETS?

Is the PCP responsible for these results?

Dataset	MaxEnt Accuracy *
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

PCP label was added to MacMorpho (v1) tagset precisely to avoid the endless discussion among (human) annotators about whether past participles should be annotated as verbs or adjectives.

Comparison of confusion between the scenarios with and without PCP.

Golden PoS	Predicted PoS	Confusion	
		with PCP	without PCP
VERB	NOUN	188	261
NOUN	VERB	182	271
ADJ	VERB	**53**	**295**
VERB	ADJ	**32**	**287**

TEST 2 – DOES LEARNING INCREASE WITH A BIGGER TRAINING SET?

A) Training with **MacMorpho-UD**; evaluation with **Bosque-UD test**;

B) Training with the **Bosque-UD train**; evaluation with **Bosque-UD test**;

C) Training with **MacMorpho-UD**; evaluation with **Bosque-UD complete corpus**

TEST 2 – DOES LEARNING INCREASE WITH A BIGGER TRAINING SET?

A) Training with **MacMorpho-UD**; evaluation with **Bosque-UD test**;

B) Training with the **Bosque-UD train**; evaluation with **Bosque-UD test**;

C) Training with **MacMorpho-UD**; evaluation with **Bosque-UD complete corpus**

Scenario	MaxEnt Accuracy
A	0.7762
B	0.9504
C	0.7647

TEST 2 – DOES LEARNING INCREASE WITH A BIGGER TRAINING SET?

A) Training with **MacMorpho-UD**; evaluation with **Bosque-UD test**;

B) Training with the **Bosque-UD train**; evaluation with **Bosque-UD test**;

C) Training with **MacMorpho-UD**; evaluation with **Bosque-UD complete corpus**

??????

Differences between training and test material???

...Both corpora contain only newspaper texts!

Bosque-UD

Scenario	MaxEnt Accuracy
A	0.7762
B	0.9504
C	0.7647

MacMorpho-UD	0.9607

...Test 1

TEST 2 – DOES LEARNING INCREASE WITH A BIGGER TRAINING SET?

A) Training with **MacMorpho-UD**; evaluation with **Bosque-UD test**;

B) Training with the **Bosque-UD train**; evaluation with **Bosque-UD test**;

C) Training with **MacMorpho-UD**; evaluation with **Bosque-UD complete corpus**

Scenario	MaxEnt Accuracy
A	0.7762
B	0.9504
C	0.7647

Bosque-UD

MacMorpho-UD	0.9607

...Test 1

??????

Differences in training and test material???
...Both corpora contain only newspaper texts!

Bezark et al. (2016) "Anchoring and Agreement in Syntactic Annotations"

- "parser bias": when (manually) revising a corpus, there is a bias that favours the parser analysis.

→ When revising a corpus, there is a bias that favours the previous analysis

CONCLUDING REMARKS

Synergy between linguistics and computer science

- Experiment 1: "What are the linguistic distinctions most effective/suitable to NLP understanding?" (??)
  - Resources to foster this kind of investigation
  - Is UD+PCP the best tagset?
    - Hmmm….. UD+PCP just postpones the problem - from the pos level to the syntactic level
  - Linguistic observation: PoS classes do not convey "natural" classes, but "consensual" ones instead. At the same time the linguistic annotation method forces us to use clear categorizations, such as traditional pos classes, it also shows us how language resists to this practice.

Experiment 2: Variation in size may not be significant (?)

- … don't forget the "parser" bias !

CONCLUDING REMARKS

Contribution:

- 3 Datasets: MM-UD corpus; MM-UD + PCP; revised version of MacMorpho v.1
- Conversion and alignment rules for reproduction of the experiments with versions 2 and 3 of MacMorpho

REFERENCES

ALUISIO, Sandra; PELIZZONI, Jorge; MARCHI, Ana Raquel; DE OLIVEIRA, Lucélia; MANENTI, Regiana; MARQUIAFAVEL, Vanessa. ***An account of the challenge of tagging a reference corpus for brazilian portuguese***. In: Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language. PROPOR. (2003)

BERZAK, Yevgeni, Yan HUANG, Andrei BARBU, Anna KORHONEN & Boris KATZ. ***Anchoring and Agreement in Syntactic Annotations***. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016).

KILGARRIFF, Adam; KOSEM, Iztok. **Corpus tools for lexicographers**. In: Granger, S., Paquot, M. (eds.) Electronic Lexicography, chap. 3. Oxford University Press (2012).

MUNIZ, H., CHALUB, F., RADEMAKER, A.: ***Cl-conllu: dependências universais em common lisp***. In: V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic). Uberlândia, MG, Brazil (2017)

NIVRE, J., de MARNE e, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: ***Universal dependencies v1: A multilingual treebank collection***. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).

RADEMAKER, A., CHALUB, F., REAL, L., FREITAS, C., BICK, E., de PAIVA. ***Universal dependencies for portuguese***. In: Proceedings of the International Conference on Dependency Linguistics. Pisa, Italy (Sep 2017)

TRUGO, L.F.: ***Classes de palavras - da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets***. Master's thesis, PUC-Rio (Aug 2016)

Thank you!

claudiafreitas@puc-rio.br

alexrad@br.ibm.com