Tagsets and Datasets: Some Experiments Based on Portuguese Language

Cláudia Freitas[1], Luiza F. Trugo[1], Fabricio Chalub[2], Guilherme Paulino-Passos[2,3], and Alexandre Rademaker[2,4](✉)

[1] PUC-Rio, Rio de Janeiro, Brazil
claudiafreitas@puc-rio.br, luizafrizzo@gmail.com
[2] IBM Research, Rio de Janeiro, Brazil
{fchalub,gpaulino,alexrad}@br.ibm.com
[3] UFRJ/COPPE/PESC, Rio de Janeiro, Brazil
[4] FGV/EMAp, Rio de Janeiro, Brazil

Abstract. We report the results of two experiments aimed at investigating the impact of linguistic variation on *PoS* tagging. In both cases, we depart from the conversion of the corpus MacMorpho [1], which was re-annotated according to the Universal Dependencies *PoS* tagset. Throughout the conversion process, we faced some linguistic challenges related to the past participle forms. As a result, we created two corpora (MacMoprho-UD and MacMorpho-UD+PCP). We used these three corpora (MacMorpho; MacMoprho-UD and MacMorpho-UD+PCP) to assess the impact on *PoS* learning in different scenarios.

Keywords: Corpus annotation · Tagset alignment · Past participles

1 Introduction

Part-of-speech (*PoS*) tagging is one of the earliest steps of NLP. However, in spite of its linguistic nature, linguistic studies comparing the impact of different tagsets on the performance of NLP systems are scarce – from the early work of [11], which examines the grammatical constructions which cause statistical n-grams taggers to falter more frequently, and the Eagles document produced in 1996, this seems to be an unattractive subject. On the other hand, comparison of PoS-taggers doesn't suffer from the same problem. A possible reason for this 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. Additionally, a requirement for this kind of study is the existence of comparable materials – the same corpus annotated with different tagsets – a requirement that seems wasteful considering the amount of work involved.

For the Portuguese language, the disparity between studies focusing on taggers and on tagsets is the same. [13] assess the ambiguity in *PoS* tagging, considering a morphologic parser. Twenty years later, [7] is the only work we know

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 459–469, 2018.
https://doi.org/10.1007/978-3-319-99722-3_46

which focuses on tagsets, and it is conducted from the computer science perspective only. As to PoS tagger evaluation, on the other hand, the scenario is far richer, as seen in [9].

In this paper, we present a study on tagsets, originated from the conversion of the corpus MacMorpho [1], which was re-annotated, at the *PoS* level, according to the Universal Dependencies (UD) tagset [16]. Throughout the process, we faced some linguistic challenges, especially with past participle forms. As a result, we created a second version of the corpus, in which we kept the UD tagset but added the PCP tag - specific for past participles – used in the original MacMorpho corpus/tagset. We then conducted two experiments: the first one aimed to verify the impact of tagsets – original MacMorpho; MacMorpho-UD; and MacMorpho-UD+PCP – on system performance; in the second experiment, taking advantage of the converted Mac-Morpho-UD corpus, we assessed the impact of size and quality in training: we used both UD-Portuguese-Bosque [18] and MacMorpho-UD in the training phase.

2 Corpus Conversion

MacMorpho [1] is a Brazilian Portuguese newswire corpus developed within the scope of the Lacio-Web project [2]. It contains 1.1 million words, annotated at the *PoS* level and manually revised. The tagset uses 23 labels. Since 2003, MacMorpho underwent two revisions, eliminating noise and making changes in the tagset, thus generating MacMorpho version 2 [7] and version 3 [6]. In the first review (version 2), repeated sentences and sentences with missing words were eliminated. There were also tokenization changes – contractions were considered a single token. In the second revision (version 3), more repeated sentences or sentences with missing words were withdrawn, and some PoS tags were simplified, aiming at achieving a coarser tagset: "auxiliary verb", "relative connective pronoun" and "relative connective adverb" were annotated with the more general labels "verb", "connective pronoun" and "connective adverb", respectively. For the present venture, we used the original version of MacMorpho (version 1), annotated with the original tagset. The choice was motivated by practical issues regarding the conversion: the Universal Dependencies scheme undo contractions and contains the tag "auxiliary verb" (AUX). Nevertheless, MacMorpho tagset is more granular than the UD tagset, with differences that are not easily circumvented with automatic alignment, thus making the alignment task a source of interesting linguistic challenges, the main one being *past participle* forms (for a detailed version of the conversion, see [22]).

2.1 The Target: Universal Dependencies Tagset

Focusing on multilingual NLP, Universal Dependencies (UD) [16] is a framework for cross-linguistic grammatical annotation that aims at developing a language-independent annotation scheme, flexible for specific extensions of a given language.

The initial UD *PoS* tasgset was proposed by [17], and consisted of 12 part-of-speech categories. The current UD tagset contains 17 labels, plus specific features that can be used to achieve a more fine-grained classification.

The UD 2.0 release contains two corpora for the Portuguese language, and one of these is the UD-Portuguese-Bosque [18]. This corpus, however, is relatively small, with 9,370 sentences and 244,675 words. Although missing (to date) syntactic dependencies, MacMorpho corpus has also been revised, and it is widely used in Portuguese *PoS* training. Besides making more material available in a context that looks promising for cross-lingual NLP, we could also investigate the impact of different tagsets in NLP.

2.2 The Alignment

The alignment of different tagsets is not a purely mechanical task of conversion. The same label can be used for different purposes, and the fact that two corpora are annotated with the same tagset does not guarantee that they are aligned[1]. For example, MacMorpho's tagset has the label `NUM`, used for numerals. The UD tagset has exactly the same tag, standing for the same category. However, while in the UD scheme the orientation is that cardinal numbers should be marked with `NUM`, in the MacMorpho corpus, if the numeral is functioning as the head of a NP, it should be tagged as `NOUN`. The MacMorpho label `PCP` has no direct equivalent in the UD scheme, and along the conversion process, we must choose between VERB or ADJ. In the following section, we specifically address the participles.

Past Participles. In modern Portuguese grammars, there is no doubt that participle forms integrate the class of verbs. However, when we look at the history of grammatical thought, we came across some interesting facts. In the *Téchné grammatiké*, which conveys some seminal ideas of what we mean by grammar, they are an independent class [3]. For the Stoic philosophers (301 B.C.), participles were considered *verbal names*, verbs with cases, among other terms that show their hybrid nature. 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: nouns and verbs [14]. There is a large number of occasions in which participles present syntactic properties of both adjectives and verbs, making a clear-cut identification nearly impossible (See [22] for a comparison of contemporary Portuguese grammars regarding past participle forms). Sentence 1 bellow illustrate this point:

(1) Refiro-me mesmo àqueles programas de interesse mais geral, como as telenovelas, já que nem essas entram às horas *anunciadas*. (I'm referring to those programs of more general interest, such as soap operas, since neither do they start at the announced/expected hours.)

[1] In this paper, we use the term *alignment* in a broad sense, meaning being equated.

This is not a specific feature of the Portuguese language, as indicated by [10]. In order to validate the linguistic decisions underlying the PCP conversion, we conducted a linguistic experiment aimed at verifying the classification of past participle forms by professionals with solid linguistic background. Not surprisingly, the results showed a huge divergence in classifications. In fact, there were sentences in which half of the volunteers used the VERB label and the other half used the ADJ label. Details of the experiment are presented in [22]. Taking into account the difficulty in deciding how to distribute past participles into VERB or ADJ, we decided to create a second corpus, with a hybrid tagset: UD labels plus PCP[2]. It is worth noting that the PCP label was added to the MacMorpho tagset precisely to avoid the endless discussion among (human) annotators about whether past participles should be annotated as verbs or adjectives.

3 Setting Up the Scene

Corpus and Tagset Conversion. To perform the tagset conversion, we created a set of general rules and a set of specific rules, designed to account for individual cases. When there were directly equivalent tags in the tagsets, the task was simple: we wrote a rule to convert all occurrences of a tag (e.g. PREP) to its correspondent in the other tagset (in this case, ADP). When there was no direct equivalence, the procedure was significantly more laborious. To convert PCP, for example, we first took a sample of 200 cases where the label appeared and read the sentences looking for patterns that could become general conversion rules.

A library in Common Lisp was developed in order to apply the rules. This library is freely available in https://github.com/own-pt/cl-tag-rewriting and it is already incorporated in the CL-CONLLU library [15]. The library produces not only the output data but also some detailed report of the rules applied to each sentence (log files). We have also developed auxiliary functions in the library to analyze the log files producing some statistics about the rules applications that helped us, for instance, to identify superfluous rules.

Since the rules are stored independently of the corpus, it is trivial to recreate the corpus. So, even though we used MacMorpho v1 for conversion, converting the material based on v2 (where sentences were deleted) or v3 is automatic (although in the latter case some minor work on merging labels will be required). In this way, we also make it feasible to study the impact on predictive models of eliminating noise from data, a point that has not yet received the relevance it deserves [19].

In order to perform the two experiments described in the following section, we run the Maximum Entropy model (Generalized Iterative Scaling method), provided by the OpenNLP suite. For both experiments, we merged *train* and *development* partitions.

[2] The UD+PCP corpus, as well as the hybrid tagset, was created with the purpose of serving as a basis for linguistic and computational experiments; and it is not our intention to integrate it into the UD consortium.

4 Experiments

The result of this conversion process is the genesis of three corpora: MacMorpho-UD (MacMorpho corpus annotated with the UD tagset), MacMorpho UD+PCP (MacMorpho corpus annotated with the UD tagset plus the PCP label) and the original MacMorpho.

4.1 First Experiment

We used the Maximum Entropy (MaxEnt) model provided by the OpenNLP suite. For each dataset (UD, UD+PCP and original), we trained in the train+dev partitions and evaluated in the test partition, as provided in the MacMorpho website. Table 1 presents the results according to each tagset.

Table 1. Learning results considering each corpus/tagset

Dataset	MaxEnt accuracy
MacMorpho-UD+PCP	0.9624
MacMorpho-UD	0.9607
MacMorpho	0.9594

The UD+PCP scenario obtained the best results, with a slightly higher performance than the UD tagset. In general, the results point to the success of less granular tagsets, but they also indicate that the criterion of granularity is not absolute: the slight advantage of UD+PCP on UD suggests that, in certain cases, the creation of a class can be a facilitator in learning, bringing consistency. On the other hand, we know that the creation of the PCP label, within the scope of the Lacio-Web project, was due precisely to the lack of agreement in the annotation of past participle forms. Thus, the results suggest that, regarding consistency in learning PoS, the creation of the PCP tag seems to have been an appropriate decision.

In order to verify whether the difference between UD and UD+PCP was due to the ambiguity of the past participle forms, we did an error analysis, starting from the confusion matrix of each scenario. Tables 3, 4 and 5 present the confusion matrix of UD+PCP and UD, respectively (predicted labels are the lines; golden labels are the columns).

As to the UD+PCP scenario, the pairs AUX-VERB (18%), PRON-DET (7%) and ADJ-NOUN (5%) are responsible for most of the confusion. The first error type/confusion (AUX-VERB) comes as no surprise. Auxiliary verbs are also verbs, and the definition of which verbal constructions should be considered auxiliaries may vary not only between languages, but also between grammarians in the same language (compare, for example, the analyses provided by [4,5,20]). It is not a coincidence that one of the changes that took place from version 2 to version 3 in the MacMorpho corpus was the elimination of the distinction between auxiliaries and verbs. The confusion between DET and PRON corresponds to ambiguous words such as *o*, *a*, *os*, *as*, *todos*, *todas*, that can be either PRON or DET, depending on the context.

Finally, the confusion between ADJ and NOUN is an old acquaintance of all students of Linguistics – in this case, we speak of a certain “fluctuation” between nouns and adjectives, which is mainly due to the common practice of naming something from its qualifications.

It is also interesting to note the confusion between PCP and NOUN (almost 4%). There is a vast amount of nouns in Portuguese which result from a lexicalization of participle forms, such as *resultado*. The analysis of a sample of this confusion suggests that this was the case.

However, to elucidate the difference between the tagsets, the most interesting confusion is the one between ADJ-VERB and ADJ-PCP on one hand, and VERB-NOUN and VERB-PCP on the other. None of the four cases stands out with the tagset UD+PCP.

Considering the confusion matrix in the MacMorpho-UD scenario, as expected, the main points of confusion observed in the UD+PCP tagset remained. However, interestingly enough, the confusion matrix also showed a growth in confusion between ADJ, VERB and NOUN. Table 2 compares the classes VERB, NOUN and ADJ in the scenarios with and without the PCP label. The results indicate that the best performance in the tagset UD+PCP is actually the result of the addition of this label, which plays the role of a disambiguator, artificially constructing a consensus where there is none.

Table 2. 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**

4.2 Second Experiment

In a second experiment, we used the MacMorpho-UD corpus to verify the impact of both training size and quality on learning. We created the following test scenarios (in all of them we run the Maximum Entropy model used in Experiment 1):

(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 Bosque-UD).

Table 3. Confusion matrix for Experiment 1, dataset MacMorpho-UD.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	8869	18	61	0	0	22	0	463	12	0	3	86	0	2	0	287	0
ADP	19	32040	129	6	13	101	0	88	4	0	45	294	0	136	0	40	1
ADV	43	42	5204	1	64	86	4	39	0	0	11	31	0	50	0	13	1
AUX	3	0	0	2737	1	1	0	12	0	0	0	13	0	0	0	298	0
CCONJ	1	1	19	0	4742	0	4	3	0	0	1	90	0	6	0	0	0
DET	21	284	72	1	4	30099	1	25	109	0	249	154	0	13	1	10	0
INTJ	0	0	0	0	1	1	7	0	1	0	0	1	0	0	0	1	0
NOUN	510	98	86	6	0	7	3	39759	41	0	17	654	8	14	0	261	0
NUM	10	2	0	0	0	8	0	86	4671	0	3	50	10	0	1	0	0
PART	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PRON	5	2	11	0	2	68	0	15	1	0	3050	28	0	148	0	4	0
PROPN	86	54	16	3	9	7	1	706	34	0	8	20058	1	2	1	30	3
PUNCT	26	4	0	0	2	7	0	27	3	0	6	45	29513	2	1	4	0
SCONJ	2	29	39	0	17	16	0	4	2	0	81	18	0	4588	0	7	0
SYM	0	2	0	0	0	0	0	1	0	0	0	1	1	0	813	0	0
VERB	295	6	10	615	7	10	0	271	6	0	1	80	0	3	0	19228	1
X	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	31

Table 4. Confusion matrix for Experiment 1, dataset MacMorpho-UD+PCP.

	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PART	PCP	PRON	PROPN	PUNCT	SCONJ	SYM	VERB	X
ADJ	8123	17	59	0	0	20	0	430	14	0	31	2	86	0	2	0	32	0
ADP	21	32048	131	7	13	103	0	87	5	0	1	45	300	0	134	0	37	1
ADV	39	45	5207	1	65	87	4	39	0	0	1	12	32	0	51	0	13	1
AUX	4	0	0	2709	1	1	0	12	0	0	1	0	14	0	0	0	276	0
CCONJ	1	1	19	0	4741	0	4	3	0	0	0	1	92	0	6	0	0	0
DET	19	271	69	1	4	30100	1	24	109	0	7	249	155	0	12	1	8	0
INTJ	0	0	0	0	1	1	7	1	1	0	0	0	1	0	0	0	1	0
NOUN	497	95	85	7	0	7	3	39803	39	0	160	17	659	7	15	0	188	0
NUM	11	2	0	0	0	9	0	85	4671	0	0	3	52	10	0	1	0	0
PART	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PCP	59	2	3	3	0	4	0	83	0	0	3927	0	6	0	0	0	11	0
PRON	3	2	11	0	2	67	0	15	1	0	0	3052	28	0	143	0	2	0
PROPN	83	56	16	3	9	7	1	707	34	0	11	8	20041	1	2	1	32	3
PUNCT	25	4	0	0	2	7	0	26	3	0	3	6	46	29514	4	0	3	0
SCONJ	3	30	39	0	18	16	0	3	2	0	2	79	19	0	4594	0	6	0
SYM	0	2	0	0	0	0	0	1	0	0	0	0	1	1	0	814	0	0
VERB	53	7	7	601	6	4	0	182	5	0	17	1	69	0	1	0	16400	1
X	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	31

Table 5. Confusion matrix for Experiment 1, dataset MacMorpho.

	ADJ	ADV	ADV-KS	ADV-KS-REL	ART	CUR	IN	KC	KS	N	NIL	NPROP	NUM	PCP	PDEN	PREP	PRO-KS	PRO-KS-REL	PROADJ	PROPESS	PROSUB	V	VAUX
ADJ	0	0	8134	56	0	0	4	0	0	0	1	431	0	97	0	29	2	19	16	0	0	0	4
ADV	0	0	40	4376	1	6	1	0	4	71	38	42	0	28	0	1	59	55	87	2	0	1	13
ADV-KS	0	0	0	0	22	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ADV-KS-REL	0	0	0	3	3	120	0	0	0	0	11	0	0	0	0	0	0	0	0	0	1	0	0
ART	0	0	8	26	1	0	26793	1	0	0	4	27	0	129	100	11	3	251	4	0	0	82	66
CUR	0	0	0	0	0	0	0	484	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
IN	0	0	0	0	0	0	0	0	7	1	0	1	0	2	0	0	0	0	0	0	0	0	0
KC	0	0	4	21	0	0	0	0	4	4769	12	7	0	90	1	0	7	2	0	1	0	2	3
KS	0	0	1	33	20	20	0	0	0	10	1962	6	0	10	1	1	14	23	7	16	105	15	5
N	1	0	531	102	0	0	2	0	3	0	10	41533	1	714	104	152	3	72	4	3	2	1	12
NIL	0	0	0	0	0	0	0	0	0	0	0	2	141	2	0	0	0	0	0	0	0	0	0
NPROP	2	0	82	12	0	0	6	0	1	11	2	729	11	19885	4	11	0	47	1	0	0	1	6
NUM	0	0	3	1	0	0	5	0	0	0	0	103	0	21	3159	0	0	0	0	0	0	0	2
PCP	0	0	60	4	0	0	6	0	0	0	0	75	0	4	0	3938	0	1	0	0	0	0	2
PDEN	0	0	0	43	0	0	0	0	0	10	3	3	0	2	0	0	1062	10	2	0	0	0	2
PREP	0	0	18	148	31	10	98	0	0	20	95	90	0	313	2	1	36	31780	9	0	0	11	35
PRO-KS	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	202	28	0	36
PRO-KS-REL	0	0	0	0	0	5	3	0	0	0	152	0	0	1	0	0	1	2	5	69	1834	0	17
PROADJ	0	0	17	45	0	0	0	0	1	1	3	8	0	10	0	2	5	1	3265	7	3	1	75
PROPESS	0	0	1	0	0	0	3	0	0	1	54	4	0	17	0	0	1	0	1	0	2	2105	2
PROSUB	0	0	2	10	0	0	22	0	0	1	4	10	1	8	0	1	1	3	38	76	8	1	959
V	0	0	44	13	0	0	1	0	0	3	0	174	1	64	0	12	14	6	5	3	0	0	2
VAUX	0	0	3	0	0	0	0	0	0	2	0	8	0	12	0	1	1	0	0	1	0	0	0

This gives us the following three issues for analysis, and the results are in Table 6: 1. The impact of training size on learning, through the comparison between scenario B and experiment 1 (corpus MacMorpho UD). 2. The impact of variation on training and test material, by comparing scenarios A and B. We note that both corpora (MacMorpho and Bosque) are composed by newspaper texts, therefore we do not expect much variation. 3. The variation in the size of the test material, by comparing scenarios A and C.

Surprisingly (for us), the results are mostly affected by the quality of the training and test materials. The surprise comes from the fact that both corpora (MacMorpho and Bosque) are journalistic texts, with overlapping materials.

Table 6. Results for scenario

Scenario	MaxEnt accuracy
A	0.7762
B	0.9504
C	0.7647

The fact that the performance of the model with the MacMorpho-UD corpus was good in experiment 1 (0.9607) and close to that of scenario B (0.9504) discards the hypothesis that the weaker performance in scenarios A and C results from inconsistencies in the annotation of the MacMorpho-UD. In other words: in both experiments 1 and scenario B, we trained and tested with the same corpus (MM-UD and Bosque-UD, respectively) and, in both cases, the results are close and equally good. We could expect internal inconsistencies in MM-UD and Bosque-UD to lead to poor performance, but that was not the case. Furthermore, since the difference between scenario B and scenarios A and C is the variation between training and test material, the difference in results suggests one of the two possibilities: (1) an alignment inconsistency between Bosque-UD and MacMorpho-UD; or (2) the interference of quality of training and test materials – in scenarios A and C the corpus used for training was different from the corpus used for testing. To investigate the alignment hypothesis, we turned again to the analysis of the confusion matrices, looking for patterns that could indicate inconsistencies between the datasets.

The main confusions are distributed in 3 groups: AUX-VERB; PRON-SCONJ and ADJ-NOUN. The confusion between ADJ-NOUN and AUX-VERB repeats the results observed in experiment 1; the only difference is in the new PRON-SCONJ confusion, which might suggest an inconsistency in alignment. Analysis of a sample of divergent cases, however, rejects this explanation. Almost all cases refer to the form *que* (*that*), which is ambiguous between relative pronoun and subordinate conjunction. The (few) remaining errors refer to *quem* (*who*) or *quantos* (*how many*). In none of the cases we noticed systematic errors from the golden corpus that suggested alignment problems.

Regarding the impact of training size on learning, and if we compare only experiment 1 (dataset MacMorho-UD) with scenario B – and assuming that both datasets are equally consistent – we observe a 1% improvement in learning when there is more training material. It is worth remembering that [8], in the context of cross-lingual parsing, indicate that the size of the training corpus ceases to be relevant from a certain amount of data.

Finally, the slightly lower results of scenario C when compared to scenario A suggest that, with more room for evaluation, performance will decline.

5 Side Effect: Optimization of Corpus Revision

Throughout the process of analyzing the confusion matrix, we were faced with confusion data that was not derived from system errors, but from the golden corpus instead (or errors from both system and golden corpus). Besides pointing out to an unfair penalization of systems, such errors served as a strategy for an optimized revision of the MacMorpho corpus annotation. All datasets used in the present study (and made available in https://github.com/own-pt/macmorpho-ud) have already incorporated such revisions.

6 Concluding Remarks

We reported here the results of two experiments aimed at investigating the impact of (i) variation on tagsets, and (ii) the size and quality of dataset training in learning.

The first experiment was empirically and theoretically motivated. The empirical motivation comes from the conversion of the *PoS* annotation of the MacMorpho corpus, which was re-annotated with the Universal Dependencies tagset. It also comes from the fact that there is a lack of an environment that enabled testing with tagsets. From a theoretical point of view, the study is linguistically motivated and takes as its starting point the well-known discussion about past participle forms. As to the second experiment, the results suggest that variation in size may not be as significant as variation in quality.

Error analysis made us realize that much of the systematic confusion reproduces the human divergence in linguistic analysis. The PCP label, the focus of the study, rightly highlights a boundary zone between two classes. At this point, it is worth remembering that the development of post-taggers arises as an engineering response to the problem of explosion due to the ambiguity of classes [21]. But the challenge remains when what is at stake is not ambiguity but “fuzziness”. This seems to be the case with past participles. This point relates to what [12] indicates as non-categorical representations, taking as an example the V-ing forms of English, which may be ambiguous between nouns and verbs in the gerund. Such cases would favor the idea of non-discrete classifications in the language, with which we agree. When the linguistic annotation method forces us to use clear categorizations, such as traditional parts of speech classes, it also shows us how limited this practice can be.

Also in regard to tagsets, an important point is to conduct the evaluation in subsequent tasks, such as syntactic dependencies. The idea is to investigate the extent to which it is important to disambiguate the confusions detected for the subsequent NLP tasks.

As additional contributions of this study, we made available to the community two corpora, plus a revised version of MacMorpho v.1, and we encourage the community to repeat our experiments. We also provide the conversion and alignment rules for easily reproduction of the experiments with versions 2 and 3 of MacMorpho. Finally, we indicate that the strategy for error analysis based on confusion matrix seems to be a good way to optimize the linguistic revision. This is a hypothesis we are investigating and in the near future we hope to develop a suite for testing and reviewing tagsets.

References

1. Aluísio, S., Pelizzoni, J., Marchi, A.R., de Oliveira, L., Manenti, R., Marquiafável, V.: An account of the challenge of tagging a reference corpus for Brazilian Portuguese. In: Mamede, N.J., Trancoso, I., Baptista, J., das Graças Volpe Nunes, M. (eds.) PROPOR 2003. LNCS (LNAI), vol. 2721, pp. 110–117. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45011-4_17
2. Aluísio, S.M., Pinheiro, G.M., Finger, M., das Graças V. Nunes, M., Tagnin, S.E.O.: The Lacio-Web project: overview and issues in Brazilian Portuguese corpora creation. In: Proceedings of Corpus Linguistics. UCREL Technical Papers (2003)
3. Auroux, S.: La révolution technologique de la grammatisation: introduction à l'histoire des sciences du langage. Philosophie et langage, Mardaga (1994)
4. Bechara, E.: Moderna Gramática Portuguesa. Nova Fronteira (2012)
5. Cunha, C., Cintra, L.: Nova gramática do português contemporâneo. Obras de referência, Lexikon (2008)
6. Fonseca, E.R., G Rosa, J.L., Aluísio, S.M.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. J. Braz. Comput. Soc. **21**(1), 2 (2015). https://doi.org/10.1186/s13173-014-0020-x
7. Fonseca, E.R., Rosa, J.L.G.: Mac-Morpho revisited: towards robust part-of-speech tagging. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (2013)
8. García, M., Gamallo, P.: A rule-based system for cross-lingual parsing of romance languages with universal dependencies. In: CoNLL Shared Task (2017)
9. García, M., Gamallo, P., Gayo, I., Cruz, M.A.P.: PoS-tagging the Web in Portuguese. National varieties, text typologies and spelling systems. Procesamiento del Lenguaje Nat. **53**, 95–101 (2014)
10. Kilgarriff, A., Kosem, I.: Corpus tools for lexicographers. In: Granger, S., Paquot, M. (eds.) Electronic Lexicography, Chap. 3. Oxford University Press (2012)
11. Macklovitch, E.: Where the tagger falters. In: Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation, pp. 113–126 (1992)
12. Manning, C.D.: Computational linguistics and deep learning. Comput. Linguist. **41**(4), 701–707 (2015). https://doi.org/10.1162/COLI_a_00239
13. Medeiros, J.C., Marques, R., Santos, D.: Português Quantitativo. In: Actas do 1° Encontro de Processamento da Língua Portuguesa (escrita e falada) EPLP 1993, pp. 33–38, 25–26 de Fevereiro 1993
14. de Moura Neves, M.: A vertente grega da gramática tradicional uma visão do pensamento grego sobre a linguagem. UNESP (2005)
15. 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). https://sites.google.com/view/tilic2017/

16. Nivre, J., et al.: Universal dependencies v1: a multilingual treebank collection. In: Calzolari, N., et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France, May 2016
17. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. In: Chair, N.C.C., et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012). European Language Resources Association (ELRA), Istanbul, Turkey, May 2012
18. Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., de Paiva, V.: Universal dependencies for Portuguese. In: Proceedings of the International Conference on Dependency Linguistics. Pisa, Italy, September 2017
19. Redman, T.C.: If your data is bad, your machine learning tools are useless. Harvard Business Review, 02 April 2018. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless
20. da Rocha Limax, C.H.: Gramática Normativa Da Língua Portuguesa. José Olympio (2010)
21. Santos, D.: POS tagging: clarificaão histórico-terminológica, 29 de Junho - 3 de Julho 2009. http://www.linguateca.pt/Diana/download/SantosEdV2009PoS.pdf
22. Trugo, L.F.: Classes de palavras - da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets. Master's thesis, PUC-Rio, August 2016