Proceedings of

LexSem+Logics 2016

The First Workshop on **Lexical Semantics for Lesser-Resourced Languages**
and
The Third Workshop on **Logics and Ontologies**

Wednesday 13th July, 2016
Tomar, Portugal

Co-located with

PROPOR 2016

The 12th International Conference on
the Computational Processing of the Portuguese Language:

Introduction

Lexical semantics continues to play an important role in driving research directions in NLP, with the recognition and understanding of context becoming increasingly important in delivering successful outcomes in NLP tasks. Besides traditional processing areas such as word sense and named entity disambiguation, the creation and maintenance of dictionaries, annotated corpora and resources have become cornerstones of lexical semantics research and produced a wealth of contextual information that NLP processes can exploit. New efforts both to link and construct from scratch such information – as Linked Open Data or by way of formal tools coming from logic, ontologies and automated reasoning – have increased the interoperability and accessibility of resources for lexical and computational semantics, even in those languages for which they have previously been limited.

LexSem+Logics 2016 combines the 1st Workshop on Lexical Semantics for Lesser-Resources Languages and the 3rd Workshop on Logics and Ontologies. The accepted papers in our program cover a number of topics across these two areas, including: the encoding of plurals in Wordnets, the creation of a thesaurus from multiple sources based on semantic similarity metrics, and the use of cross-lingual treebanks and annotations for universal part-of-speech tagging. We also have talks from two distinguished speakers: on Portuguese lexical knowledge bases (different approaches, results and their application in NLP tasks) and on new strategies for open information extraction (the capture of verb-based propositions from massive text corpora).

We would like to take this opportunity to thank you for your involvement and participation in LexSem+Logics 2016, and hope that you enjoy the workshop!

The Organizers

Organizers

Steven Neale – University of Lisbon, Portugal
Valeria de Paiva – Nuance Communications, USA
Arantxa Otegi – University of the Basque Country, Spain
Alexandre Rademaker – IBM Research & FGV/EMAp, Brazil

Program Committee

Cláudia Freitas – PUC-Rio, Brazil
Nuno Mamede – INESC-ID, Portugal
Gerard de Melo – Tsinghua University, China
Hugo Gonçalo Oliveira – University of Coimbra, Portugal
Petya Osenova – Bulgarian Academy of the Sciences, Bulgaria
Thiago Pardo – University of São Paulo (São Carlos), Brazil
Adam Pease – Articulate Software, USA
Livy Real – IBM Research, Brazil
Kiril Simov – Bulgarian Academy of the Sciences, Bulgaria

Invited Speakers

Hugo Gonçalo Oliveira – University of Coimbra, Portugal
Pablo Gamallo – University of Santiago de Compostela, Spain

Workshop Program

12:00pm – 12:30pm	Invited Talk: Hugo Gonçalo Oliveira *Portuguese Lexical Knowledge Bases*
12:30pm – 13:30pm	Lunch
13:30pm – 14:00pm	Livy Real & Valeria de Paiva *Plurality in Wordnets*
14:00pm – 14:30pm	Filipe Islaji de Albuquerque & Hugo Gonçalo Oliveira *Dicionário Creativo: The Construction of a Fuzzy Onomasiological Thesaurus from Multiple Sources*
14:30pm – 15:00pm	Valeria de Paiva & Livy Real *Universal POS Tagging for Portuguese: Issues and Opportunities*
15:00pm – 16:00pm	Invited Talk: Pablo Gamallo *Strategies for Open Information Extraction*
16:30pm – 17:00pm	Coffee Break (in Lobby Bar)

Contents

Invited Talks – Abstracts

Invited Talks – Abstracts

Hugo Gonçalo Oliveira
University of Coimbra, Portugal

Portuguese Lexical Knowledge Bases

Lexical knowledge bases (LKBs) are resources structured in words and their meanings, typically used as dictionaries for computers. While, for English, Princeton WordNet became the paradigmatic resource of this kind, for Portuguese, there are several alternatives, developed by different teams, with different approaches and long-term goals, which resulted in different resources with different strengths and limitations. But none is as consensual as Princeton WordNet.

This talk will make a journey through existing LKBs for the computational processing of Portuguese. Several resources of this kind will be presented together with their availability, creation approach, size, current limitations and future goals, among other features. Moreover, besides a shallow qualitative comparison, natural language processing tasks where these resources could and have been useful will also be presented.

Pablo Gamallo
University of Santiago de Compostela, Spain

Strategies for Open Information Extraction

Open Information Extraction (OIE) is an emerging field in Information Extraction interested in applying shallow semantics techniques and unsupervised learning methods to extract great amounts of basic propositions (verb-based triples) from massive text corpora which scales to Web-size document collections.

An OIE system reads in sentences and rapidly extracts one or more textual assertions, consisting in a verb relation and two arguments, which try to capture the main relationships in each sentence. We will introduce the main properties of this extraction method as well as the different types of strategies proposed so far.

Plurality in Wordnets

Livy Real and Valeria de Paiva

[1] IBM Research, Brazil
[2] Nuance Communications, USA
livym@br.ibm.com valeria.depaiva@nuance.com

Abstract. We investigate the features of Princeton WordNet associated with nouns that are essentially plural. This means exploring the Princeton WordNet feature `classifiedByUsage: plural` that labels synsets and words commonly used in the plural. We decided to investigate how this feature works for Portuguese and here we discuss the best way to encode this kind of lexical information in OpenWordnet-PT, an open wordnet for Portuguese.

1 Introduction

Lexical resources are required for several Natural Language Processing tasks. They are canonically used on word disambiguation tasks [1], in information retrieval [6], as input to several ontologies [7], and more recently as features when building space vectors for machine learning approaches [5] or in anaphora resolution [8]. What we commonly expect from lexical resources is that they should collect all lexical information one needs for processing texts, from syntactical frames to semantic features. Wordnets are one of the most used lexical resources and they provide generally syntactic, semantic and even contextual/pragmatical information about words and their senses. Here we focus on how wordnets encode a lexical plurality feature. This is an idiosyncratic feature that tell us that a given word is often used in the plural, as for example the English words *glasses* and *manners*.

We start by looking at how Princeton Wordnet (PWN) encodes this information, through a pointer `ClassifiedByUsage` applied to a specific synset domain `plural`. Then we look at an ongoing wordnet project for the Portuguese language, the OpenWordnet-PT (OWN-PT). The Portuguese lexicon OWN-PT is a projection of PWN and has all PWN relations and features, hence OWN-PT inherits the plural classification from PWN. However, as it is expected, plurality does not hold in Portuguese for the same word senses that it does in English. The word for *glasses*, for example, in Portuguese *óculos*, can be used both in the singular or in the plural having the same real world reference, a pair of glasses.

The issue of whether objects are referred to in the plural or not is related to the question of how to encode information about mass nouns, that is, whether the countable/uncountable distinction, which distinguishes words such as *apple* (countable) and *blood* (uncountable) needs to be marked on the lexicon and how. As preparation for deciding on the best way to have the countable vs.

uncountable lexical information encoded in OpenWordnet-PT, we investigate how the PWN plurality feature fits in with Portuguese words and how this information can be useful to Natural Language Processing (NLP) tasks.

In the next section we introduce the plurality issue, from the theoretical semantics and the NLP perspectives, pointing out some tasks where lexical information on plurality is relevant. In section 3, we recall Princeton Wordnet and its plurality feature. In section 4, we briefly introduce OpenWordnet-PT and discuss how this feature appears in Portuguese. We also present some data and statistics on how English plurality fits in with Portuguese word senses in section 4.1. To finish we discuss the best way to encode lexical plurality in OWN-PT and draw some conclusions.

2 Plurality

This section offers a brief overview of semantic studies on plurality, mainly based on Chierchia's[3] assumptions. Then we discuss how plurality has been used in NLP tasks and present our motivations for the present work.

2.1 Plurality in Semantics

Plurality has been discussed in formal linguistics studies at least since Quine's seminal work [12]. Plurality and associated issues, such as, its formalization, collective readings and the well known distinction between mass-nouns and count-nouns, have been comprehensively studied and play an important role in most theories that formalize the semantic behavior of natural language.

Here we are interested in how to code plurality information in lexical resources like wordnets. For this, we first state some definitions based on [3]. Chierchia states that what distinguishes mass nouns (as *blood, water, furniture*) from count nouns (such as *boy, drop, sofa*) is an intrinsically semantic property from which many morpho-syntactic properties follow.

Common count nouns point to what Chierchia calls *singularities* in the lexicon. They can refer both to a class of objects or to a single unit, *coin, the coin*, in sentences such as *Give me the coin* or *A coin is a piece of hard material.* Mass nouns are 'generally interpreted as a mereological whole of some kind' and the domain of its minimal components is somehow more vague than singularities, as it is the case with the word *change.* It is important to note that this is an intrinsically grammatical property. That is, this property is not related to the ontological objects those words refer to. It is the word itself that dictates whether a noun will be countable or not. Examples that show that are the pairs *coin/change*, *shoe/footwear* and *virtue/honesty.* While common count nouns express singularities, these nouns in plural express a set formed by these singularities, *boys* is a set of some individual *boy*s, in which we can still see the minimal unit, a single *boy.*

For now, we are more interested in count nouns lexicalized as plurals, such as *manners* (for example in the phrase *he has the manners of a pig*) and in nouns whose only possible form are plural, such as *pants* and *quarters*. However, we expect that our discussion on how to encode plurality in lexical resources should bring insights about how to encode information on mass/count nouns. We are also interested in collective nouns, such as *group* and *committee*, that differ from mass nouns as they can be pluralized — *groups, committees*, but not **waters, *furnitures* — and also differ from common count nouns as they already refer to sets of things.

Lexical resources, such as wordnets, in general offer only the lemma of a given word, as its dictionary form. In the case of nouns, this means the masculine and singular form, when applicable, such as *boy, manner* and *group*. Lexical resources also offer semantic relevant information, as when they group words commonly used in the plural — in PWN this feature in encoded via the feature `ClassifiedByUsage:plural` — or when a noun is a collective noun — that in PWN is described via the lexicographer files. The collective nouns are part of the file `noun.group`. A researcher interested on collective nouns or in words normally used in the plural can find enough information within PWN's state of art for English. However, the mass/count distinction is not one of PWN's classification features, at the moment.

2.2 Plurality in NLP

We briefly review the motivations for this work. First, we are interested in completing and improving the lexicon of OpenWordnet-PT. Improving lexical resources is a hard and time consuming task with no laurels, but extremely necessary for lesser resourced languages, as Portuguese still is. To ensure that OWN-PT is as informative for Portuguese as PWN is for English, and to make sure that the PWN information inherited by OWN-PT is correct, we want to check all synsets and relations in OWN-PT. However, since that is a large amount of data, we have been revising OWN-PT's content in pieces, considering different features or relations of PWN and consistently checking how these are encoded in OWN-PT.

Second, we are interested in defining notions of plurality and count-mass distinctions in lexical resources, as this information is lexical and can be used in several applied tasks. For example, knowing that some pluralized expressions actually refer to unique entities is necessary for doing textual reasoning.

Recent work of [4] points out the necessity of considering pluralized word forms when building vector space models for machine learning applications. The author notices that the distribution of words and other textual features changes when a word is often used in its plural form. Thus building word vectors only considering its lemma oversimplifies the features. However, as [11] and [14] show, using wordnet relations as features in vector space modelling can improve machine learning algorithms results. Thus it seems that this PWN classification can be helpful when modelling plurality or countability.

Another use case of this feature can be seen in [9] that proposes a new method of a picture based communication, a language independent method of communication for people with disabilities. The proposed method uses the PWN

`ClassifiedByUsage:plural` relation to improve the possibility that a given picture corresponds perfectly to a concept. Having this feature encoded allows the system to automatically recognize that a picture that contains only one element can be correctly associated to a word in a synset classified as plural. One single pants is enough to describe 'pants', differently, for example, from 'birds' that should be described by a picture containing more than one single bird.

3 Plurality on Princeton Wordnet

Princeton Wordnet(PWN) is the mother of all wordnets and has been developed for English by the Princeton team over the last three decades. PWN offers a database of nouns, verbs, adjectives and adverbs arranged into sets of cognitive synonyms (synsets). Each synset refers to a single concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. These synsets are related through many (conceptual-semantical) relations, such as synonymy, heteronomy and meronymy. Each synset has one specific ID, one or more word forms that express its sense, a gloss (a small concept definition) and many of them also have sentences exemplifying its use.

The latest version of PWN available is version 3.1, that contains some 117,000 synsets. Released in 2006, this version of PWN is still the larger and most reliable lexical resource available for English. There are many relations in PWN classified by `pointers`, that can be semantic or lexical. According to the PWN documentation, lexical pointers, such as `Antonym` and `Derivationally Related`, are *normally* used to indicate lexical relations that hold between words. Semantic pointers establish semantic relations that are *generally* used for linking synsets. Examples of semantic pointers are `Hypernym` and `Hyponym` relations, that clearly should hold between synsets.

However, pointers such as the `Region` domain, that assigns a word form to where it is usually used (in the US or England, for example) and the `USAGE` domain, that specifies content as an archaism or a plurality, are hybrid between lexical and semantic pointers, an issue that had been pointed out by Eric Kafe in the 2000's and is discussed again in [13]. The definition of PWN pointers itself is vague and PWN seems to accommodate pointers that have a hybrid range, that is, that classify different kinds of objects. However it is still an ongoing discussion on PWN and Global Wordnet Association[3] communities what to do with thes hybrid pointers.

The feature `ClassifiedByUsage:plural`, or `Domain-Usage:plural`, filters nouns synsets *and* word senses that are often used in plural. This relation applies only to nouns and labels 239 elements. Some examples are {`07942152-n people| any group of human beings (men or women or children) collectively | ``old people''`}; {`04862236-n nerves | control of your emotions | ``this kind of tension is not good for my nerves''`} and synset

[3] `http://globalwordnet.org/`.

{`04356056-n sunglasses, shades, dark glasses | spectacles that are darkened or polarized to protect the eyes from the glare of the sun| ``he was wearing a pair of mirrored shades''`}.

The PWN list of synsets classified by plural usage can be found at `http://wordnet-rdf.princeton.edu/wn31/106230167-n`. This comes from dictionaries, since this feature is idiosyncratic and there is no general pattern that can capture only those words. Although the PWN classification of plurality appears useful, the feature is not well curated. The documentation says that the word is usually used in the plural, but some of word forms appear in the plural, others in the singular and sometimes PWN have both forms in the same synset. For example we have the synset {`06630627-n regard, wish, compliments | a polite expression of desire for someone's welfare | ``give him my kind regards''; ``my best wishes''`}, where *wish* is in the singular, while *compliments* is in the plural.

4 OpenWordnet-PT

The OpenWordnet-PT (OWN-PT) is an open wordnet for Portuguese, in development since 2012 and modelled after and fully interoperable with the original PWN. The OWN-PT uses the same identifiers as the last released version of PWN and it is browsable at and downloadable from `http://wnpt.brlcloud.com/wn/`. The OWN-PT is also linked to the largest open source common sense ontology, the Suggested Upper Merged Ontology (SUMO)[4], described in [10] and to the Open Multilingual WordNet (OMW) project[5], again browsable and downloadable as described in [2]. Since the Open Multilingual Wordnet project merges dozens of wordnets, ways of improving each one of these wordnets might percolate to the other ones. Thus the plurality encoding issue discussed here can, in principle, affect and/or be useful to all of these other lexical resources.

For these reasons we would like to be sure that all the PWN relations and features that are inherited by OWN-PT are not too language specific to English and should be present in derived wordnets, such as ours. Our hypothesis was that the feature of plural usage is indeed idiosyncratic and we doubted the suitability of automatically percolating it through to the Portuguese synsets. Thus we decided to check all the synsets marked with this feature, completing all of them in OWN-PT and collecting candidates that should and should not keep this feature in Portuguese. The data derived from this methodology is investigated in the following.

4.1 Data and Statistics

From the 239 synsets that PWN marked as used in plural, OWN-PT had 72 non-empty synsets, that is, synsets with some corresponding Portuguese words added.

[4] `http://www.ontologyportal.org`

[5] `http://compling.hss.ntu.edu.sg/omw/`

Thus, our first step was to complete the empty synsets in OWN-PT, as for example, {`03684224-n locking pliers | pliers that can be locked in place`}, where the Portuguese word *alicate de pressão* was added.

From those 239 PWN synsets, we did not complete some 18 synsets, as we could not find lexicalized forms in Portuguese for those senses. For example we do not find Portuguese words to complete the synset {`04570532-n widow's weeds, weeds |a black garment (dress) worn by a widow as a sign of mourning`}. We left those synsets empty and marked them in the web interface of OWN-PT with the tag `en_only`, a tag proposed to identify PWN synsets verified, but with no translation to Portuguese. This list can be checked at `http://wnpt.brlcloud.com/wn/search-activities`, if one searches for the `en_only` hashtag.

From those 221 synsets that should have senses in Portuguese, 48 synsets have both plural and singular word forms used in Portuguese. For these we completed the synsets with the singular lemma and left the PWN plurality feature indicating that these synsets are in general used in the plural. Examples of this kind are: {`02854739-n pants, bloomers, knickers, drawers | calcinha, calça | underpants worn by women; ``she was afraid that her bloomers might have been showing"`} and {`03041449-n cleats | chuteira | shoes with leather or metal projections on the soles; ``the football players all wore cleats"`}

Around 100 synsets seem to be expressed only through singular words in Portuguese, such as {`02850552-n bleachers | arquibancada | an outdoor grandstand without a roof; patrons are exposed to the sun as linens are when they are bleached`}. We listed them as synsets that maybe should loose the PWN plurality classification, but have not changed them in our lexical base. This list is available at `https://github.com/livyreal/Singular_Synsets`.

Around 74 synsets, approximately 30% of the original PWN list, are actually often used in the plural. For those, we added the singular and the plural form in OWN-PT, as one could use this pluralized lemma information. We decided for adding also the singular form, since in many pipelines the word forms searched within wordnets are lemmas, singular forms. Synset { `07943646-n ancients | antigo, antigos | people who lived in times long past (especially during the historical period before the fall of the Roman Empire in western Europe)`} and {`00179916-n wings | asas, asa | a means of flight or ascent; ``necessity lends wings to inspiration"`} are examples of this.

Finally some 27 synsets were completed with mass nouns, which are indeed singular forms. We completed them and marked all of them with the `#mass`, labelling them for future work. Examples are {`07942152-n people | população, gente, povo | any group of human beings (men or women or children) collectively; ``old people"; ``there were at least 200 people in the audience"`} and { `02730568-n fitting, appointment | aparelhagem | furnishings and equipment (especially for a ship or hotel)`}.

4.2 Discussion

Some interesting points came up when looking this data. Besides the fact that the feature is idiosyncratic, we can sketch some preliminary conclusions. For example, nouns related to clothing and instruments, such as *pliers, tongs, pants* and *suspenders*, that are in English lexicalized as plural forms, do not have the same behavior in Portuguese: *alicate, pinça, calça, suspensório* can be used in the plural, but do not have to be so. The singular forms are perfectly acceptable. Clothing nouns in Portuguese sometimes can refer both in the singular or the plural to the same entity. The forms *calça* or *calças* can refer to a single object, but the singular usage is more frequent, at least in Brazil. The same does not hold for instruments, as *tesouras* and *alicates* only refer to more than one object. We decided to add to OWN-PT only word forms in the singular in these cases and keep the feature `ClassifiedByUsage:plural` in the synsets of words that have the same referent both in plural and in singular.

Also in the clothing domain, we found some English words that have made their way into Portuguese: *shorts, jeans*. They are interesting because they are often used in plural in English and arrived in Portuguese already with the plural mark (*-s*), but, even with the plural mark, they are common count nouns and currently used in the singular, *o meu jeans, o shorts dela*. For those, we added only the singular word forms in Portuguese.

In English, there are also many pluralized abstract nouns, as *congratulations, felicitations, compliments, regards, wishes*, that are also, in general, used in plural in Portuguese, *congratulações, felicitações, cumprimentos, parabéns, votos*. In both languages, we can find contexts where those words are used in singular: *That was a nice compliment!/Isto foi um ótimo cumprimento!*. For those, we keep the `ClassifiedByUsage:plural` feature, but add only the word form in singular, except for the case of the word *parabéns*, that has only one form.

A general remark about this `ClassifiedByUsage:plural` PWN feature is related to its vague definition that influences what kind of objects in wordnet it classifies, following remarks in [13]. The documentation of PWN says `ClassifiedByUsage:plural` labels synsets. Those synsets come with a gloss *often in plural* (or *usually plural*) and are connected to the synset {`06295235-n plural, plural form | the form of a word that is used to denote more than one`}. However, many of those synsets have more than one word form and not all of them are used in plural. This is the case of {`03504723-n central office, main office, home office, home base, headquarters | the office that serves as the administrative center of an enterprise; ‘‘many companies have their headquarters in New York"`}, in which only the word *headquarters* are actually often used in plural. Keeping this feature related to synsets would also cause some problems in Portuguese, as some synsets bring both words in plural and in singular, e.g. {`03405265-n furnishing | móvel, mobília, mobiliário | the instrumentalities (furniture and appliances and other movable accessories including curtains and rugs) that make a home (or other area) livable`}. *Móvel* only have this sense when in plural and both *mobília* and *mobiliário* are mass nouns, which show us that we need to have a wordnet story to tell about mass nouns too. For now we decided to not include in the glosses of synsets the information *usually in plural* and collecting candidates for lose this feature.

This mass/count issue also appears when looking to synsets that refer to groups, as {`08179205-n poor people, poor |people without possessions or wealth (considered as a group; ‘‘the urban poor need assistance"`} and {`08477307-n unemployed people, unemployed | people who are involuntarily out of work (considered as a group; ‘‘the long-term unemployed need assistance")`}. For those synsets, we add mass nouns (whenever is possible) and word forms in singular when we do not have a mass noun for it. We also leave those synsets labelled with the PWN plural feature. For now, we do not have a way to mark mass nouns in OWN-PT and we leave this as near future work.

5 Conclusions

This work is an investigation on how to encode plurality in lexical resources, namely we checked how Princeton Wordnet brings this information and how is the best way to fit it in OpenWordnet-PT, an open wordnet for Portuguese language.

Princeton Wordnet has the classifier feature `ClassifiedByUsage:plural`, that labels synsets and words which are often used in the plural. However one can not expect that this idiosyncratic feature could percolate to other languages. Then we checked in OpenWordnet-PT how these PWN pluralized synsets should be stated in Portuguese. Around 55% of the pluralized synsets in English are truly used in plural in Portuguese and many of them can also be used in the singular, keeping the same meaning. We then list the remains as candidates for loosing this plurality feature inherited from PWN.

From this manual checking of OpenWordnet-PT synsets, we completed more than 200 synsets in Portuguese. Even in synsets with word forms usually used in plural, we decide for completing Portuguese synsets with its singular form. We decide for this uniform treatment mostly thinking on NLP tools that search for lemmas in wordnet. Keeping their lemma (as usual in singular) and also the plural label, we think we have the correct information encoded.

However, many concepts with this plurality feature can be translated as mass nouns, which we think is the best translation in several cases. How to encode this mass feature, however, is still an open question to us, that we leave as future work.

References

1. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing. pp. 136–145. CICLing ’02, Springer-Verlag, London, UK, UK (2002), http://dl.acm.org/citation.cfm?id=647344.724142

2. Bond, F., Foster, R.: Linking and Extending an Open Multilingual Wordnet. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL, ACL, Sofia (2013)
3. Chierchia, G.: Plurality of mass nouns and the notion of 'semantic parameter'. In: Rothstein, S. (ed.) Events and Grammar, pp. p. 53–103. Kluwer (1998)
4. Katz, G., Zamparelli, R.: Meaning-shifting plurality ans the count/mass distinction. In: Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4) (2011)
5. Legrand, S., Pulido, J.: A hybrid approach to word sense disambiguation: Neural clustering with class labeling. Knowledge Discovery and Ontologies (KDO-2004) workshop, 15th European Conference on Machine Learning (ECML) and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (2004)
6. Mandala, Rila, T.T., Hozumi, T.: The use of wordnet in information retrieval. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal (1998)
7. Mann, G.: Fine-grained proper noun ontologies for question answering. In: Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks. Taipei (2002)
8. Meyer, J., ', R.D.: Using the wordnet hierarchy for associative anaphora resolution. In: Proceedings of the Coling 2002 Workshop 'SemaNet'02: Building and Using Semantic Networks. Taipei (2012)
9. Narayanan, A.: Systems and methods for picture based communication (Apr 29 2014), `https://www.google.com/patents/US8712780`, uS Patent 8,712,780
10. Niles, I., Pease, A.: Toward a Standard Upper Ontology. In: Welty, C., Smith, B. (eds.) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems. FOIS-2001 (2001)
11. Patwardhan, S., Pedersen, T.: Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL (2006)
12. Quine, W.: Word and Object. MIT Press, Cambridge (1960)
13. Rademaker, A., Chalub, F.: Verifying integrity constraints of a rdf-based wordnet. In: Global Wordnet Conference 2016. Bucharest, Romenia (Jan 2016)
14. Wibowo, A., Christian, P., Handojo, A., Halim, A.: Application of topic based vector space model with wordnet. In: Proceedings of Uncertainty Reasoning and Knowledge Engineering (URKE) (2011)

Dicionário Criativo: The Construction of a Fuzzy Onomasiological Thesaurus from Multiple Sources

Felipe Iszlaji de Albuquerque and Hugo Gonçalo Oliveira

CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Portugal.
felipeiszlaji@gmail.com, hroliv@dei.uc.pt

Abstract. This paper reports on the construction of a fuzzy onomasiological thesaurus for Dicionário Criativo, an online website specialized in creative writing. In order to build this thesaurus, we merged distinct thesauri using similarity metrics and measuring the importance of each word inside the concept represented. To shape the concepts, we first used a clustering algorithm and then a graph-based technique. This process generates larger semantic groups, the core for our thesaurus and a helpful tool for creative writing. To rank their position inside a group, we relied on the frequency of each word in each semantic group.

Keywords: thesaurus, fuzzy thesaurus, onomasiological thesaurus, synsets

1 Introduction

The process of creative writing goes beyond the bounds of technical forms of literature. The writer is responsible for composing the text by picking up words from a wide range of choices. Every word selected in this process is intended to provide the exact direction into the idea the writer wishes to convey. In order to help the creative process, some tools are suggested, namely: dictionaries, thesaurus, word lists, mind maps, among others. Despite this universe of tools, the thesaurus is probably the most helpful.

A thesaurus is a compilation of words linked to their synonyms and related concepts. In general, a thesaurus also offers a categorization system which groups words by their general meaning or semantic concept. Inside each semantic group, the words are all closely related. This categorization is crucial in the creative process. In the writer's perspective, he may know the concept he wants to express, but he may lack specific words that lexicalize it. Therefore, knowing all related words inside a semantic concept may help this core task of the creative writing process.

In opposition to definition dictionaries (semasiological dictionaries), where alphabetically sorted entries define each lexical item, thesauri first define a concept (an internalized idea) and then expands it into lexicon units. This different direction, from meaning into lexicon, is the main purpose for using thesauri in creative writing.

Thesauri are valuable for writers, including journalists, advertising copywriters, poets, novelists, songwriters and screenwriters, as well as intellectuals and academics. Information technologies have provided an increasingly accelerated digitization of information. When using the possibilities of digital technology, dictionaries and reference works may gain in functionality and usability. Moreover, the physical size limitation imposed by a printed dictionary does not exist in the digital environment. In this environment, searches by keywords, hyperlinks and the possibilities for user interaction with a graphical user interface make the access to information much easier and productive. Digital technologies providing access to information in hyperlinked ways leverages the network of morphological, syntagmatic, paradigmatic and semantic relations between different lexical units.

In this paper, we describe a related work in section 2, which includes the revision of Portuguese thesauri and Portuguese semantic networks. In section 3, we describe the multiple sources used in our proposal. In section 4, we propose an approach to create a fuzzy onomasiologial thesaurus through the combined use of two algorithms, an agglomerative clustering algorithm and a splitting algorithm. Our results are compared with the results of two other Portuguese semantic networks in section 5.

2 Related Works

The most popular English thesaurus is Roget's Thesaurus [18]. Created by Dr. Peter Mark Roget in 1805, it has been enlarged and is now available on the Web. To justify its importance, there are inumerous works that exploit this thesaurus [11]. On the other hand, WordNet [6] is the most successful digital resource of this kind. It can be seen as a thesaurus enriched by the semantic and syntagmatic relations among the words.

For Portuguese, the first published examples of onomasiological thesauri were the Dicionário Analógico da Língua Portuguesa: *ideias afins* [1] and the Dicionário Analógico da Língua Portuguesa: *tesouro de vocábulos e frases da língua portuguesa* [19]. Both works followed the ideas of Roget. Other important Portuguese thesauri are the Dicionário de Ideias Semelhantes [7] and the Dicionário de Palavras Interligadas [17]. Unfortunately, all these works are only available on printed media and some are not published anymore. For digital media, Tesauro Eletrônico do Português (TEP) [5] provides direct synonyms and antonyms, but does not provide analogies.

Not many years ago, it was common to say that Portuguese lacked lexical-semantic knowledge bases with a similar structure as WordNet. In recent years, the scenario has changed and researchers can now choose between several alternatives [10], such as WordNet.PT (WN.PT) [12] [13], WordNet.BR (WN.BR) [4], OpenWordNet-PT (OpenWN-PT) [16] and the Onto.PT [8], with which our approach has similarities.

According to its website[1], WN.PT contains a network of 10,000 concepts, including nouns, verbs and adjectives, their lexicalizations in different variants of the Portuguese and their glosses. The concepts are integrated into a network of more than 40,000 instances of relations.

As reported by [4], the full version of WN.BR will cover relations hyperonymy, part-of, cause and implication. However, this version is not available on the web. We have only the TEP numbers[2], which is the first part of the WN.BR project. It includes more than 44,000 lexical items organized in 19,888 synsets.

OpenWN-PT [10] currently has 43,925 synsets, of which 32,696 nouns, 4,675 verbs , 5,575 adjectives and 979 adverbs.

In contrast to most wordnets, Onto.PT is not handcrafted, but created automatically, by exploiting Portuguese dictionaries and thesauri. Onto.PT was released in 2012 and the last version, 0.6 (August 2013), contains 156k unique words, organised in 117k synsets, connected by 173k relation instances.

[1] http://cvc.instituto-camoes.pt/traduzir/wordnet.html

[2] http://www.nilc.icmc.usp.br/tep2/

3 Resources

Our approach uses multiple resources to build a fuzzy onomasiological thesaurus. Those include other printed thesaurus, which were integrated to form a thesaurus with higher coverage. More precisely, we used the Dicionário Analógico da Língua Portuguesa: ideias afins, TEP, Dicionário de Ideias Semelhantes and OpenThesaurus.PT. Each of them is particular and has its own structure. The first, for example, is organized as an onomasiological dictionary, with a rich system of classification, but is quite old and only available in printed media. TEP, on the order hand, is a recent thesaurus, well organized, but lacks a huge vocabulary and brings only close synonyms. Our built-in dictionary has no organization of concepts, but has an updated vocabulary and a well defined core of important synonym words. Below, we present the dictionaries used.

TEP 2.0 is an electronic dictionary with synonyms and antonyms to Brazilian Portuguese. TEP was developed by Brazilian researchers from NILC. The TEP present the entries by relevance order. It is important to note that TEP has no onomasiological information.

OpenThesaurus.PT is the Portuguese version of a collaborative thesaurus initiative [15]. Approximately four times smaller than TEP, it contains 13.258 lexical items, organised in 4.102 synsets, with 3.2 lexical units per synset, on average. The project has not had any significant development since 2006.

Azevedo's dictionary is an onomasiological dictionary for Portuguese. We used the second version, published in 2010 by Lexikon publisher. In this version, the dictionary was manually reviewed and new synsets were added by the lexicographer Paulo Geiger. This dictionary covers about 100k lexical items organized in more than 1.000 analog groups.

Dicionário de Ideias Semelhantes is an analog dictionary with words grouped by similarity. Groups are classified in nine categories: (i) sense abstract; (ii) sense affective; (iii) sense dynamic; (iv) sense physical; (v) sense moral; (vi) sense religious; and (vii) sense social.

Table 1 shows the numerical analysis of the aforementioned resources.

Table 1. Numerical analysis of the sources.

Base	#Synsets	Max(#words in synsets)	Mean(#words in synsets)
TEP	19,885	53	3.81+/-3.31
OpenThesaurus.PT	3,733	17	3.6+/-1.96
Dicionário Analógico	16,065	347	13.52+/-20.31
Ideias Semelhantes	5,424	96	8.8+/-.75

4 Methodology

The idea behind an onomasiological thesaurus is to start from the concept and move on to the lexicon -- the idea is first formalized and then we look for words that could represent it. The methodology used in this work is focused on providing a feasible and relevant automatic method for merging Portuguese thesauri. Any technique used to merge different thesauri should solve two types of problems: (i) Which synsets should be merged; and (ii) How to delimit a concept [8].

In order to illustrate these problems, we present a problem scenario. For solving (i), we have merged synsets from the different thesauri. However, this thesauri relation may not be one-to-one. An example is provided for, merging two different thesauri with the configuration below. The synsets are specified by the letter S followed by the number (S1, S2, S...). The lexicon that belongs to each group is specified by the alphabet letters. Equal letters between thesauri mean the same lexical item.

Thesaurus 1: **S1**={a,b,c}
Thesaurus 2: **S2**={a,c} **S3**={b,d,e}

In this case, we may have two intuitive possibilities. Merging S1, S2 and S3, which would result in the group {a,b,c,d,e}, or merging only S1 with S2, while leaving S3 alone, which would result in groups {a,b,c}, {b,d,e}. In this situation, the merging algorithm should have sensitive parameters to allow the right decision for most of the cases.

In this way, we proposed to first apply an agglomerative clustering algorithm to the synsets. It will start with a list of all concept groups from all thesauri. It will then rank the concept groups by similarity among their lexicon. The third step is to cluster the two groups that are most similar from the rank list. This new group will be the union of the lexicon from the two merged groups. If a lexical item belongs to both merged groups, then this concept will weight 2 in the new formed group. Then, the algorithm will remove the two merged clusters from the list and add the new group. This process is then repeated from ranking to an inclusion of the new concept groups until a threshold of minimum similarity between the groups is achieved. The algorithm stops when no more groups are above this similarity threshold.

For computing similarity, we designed a set of rules for a relative weighted measure inspired in the Jaccard similarity. These rules, presented following, have been empirically verified to be well-suited for our case. Take the intersection lexicon between the two groups. The groups are not similar if the lexicon intersection is below 2. Computed the absolute similarity by the sum of the lexicon in the intersection with their weight, return the relative weighted similarity by dividing the absolute similarity by the lower cardinality of the groups. The algorithm only accepts groups as similar if the metric returns a value above the defined threshold of 0.4. We illustrate this technique with the following example.

Thesaurus 1: **S1**={amor, afeto, afeição, carinho, simpatia}
Thesaurus 2: **S2**={amor, afeto, paixão} **S3**={amor, ardor, calor, paixão}
Thesaurus 3: **S4**={ardor, arrebatamento, calor, fervor, paixão}

In the first step we will produce a list of the concepts ranked by their similarity.

S1 ∩ **S2** = {amor, afeto} Similarity = 0.4
S1 ∩ **S3** = {amor} Similarity = 0.2
S2 ∩ **S3** = {amor, paixão} Similarity = 0.4
[...]

The best ranked groups are then merged and the lexical items weighted as shown below. A new rank taking this new group is then formulated and the process starts again.

Agglomerative algorithm output

S1 ∩ **S2** ∩ **S3** ∩ **S4** = {amor, afeto, afeição, ardor, arrebatamento, calor, carinho, fervor, paixão, simpatia}

This clustering approach leads to large concept groups. This is good because we want related groups to be merged but, on the other hand, some groups with different meaning end up being merged. To minimize this issue, we proposed a second algorithm that will address problem (ii) how to define a concept.

Although defining a concept is not an objective task, we want to have groups of words that transmit a concept in common. So, we want that the words inside a concept group to be related and to weight this relatedness for creating a fuzzy onomasiological thesaurus. For this reason, our second algorithm is a graph-based approach to be run on the output of our clustering algorithm.

This graph-based algorithm (by adjacency matrix approach) is responsible of splitting different concepts from each group formed earlier. In order to illustrate this approach, we show an example below. We show the agglomerative output in the first line, which will guide the algorithm execution. The first thing the algorithm does is to retrieve, from all the thesauri, the groups that have at least two words in common with the agglomerative output. The second step of this algorithm is to perform splits.

Table 2. Splitting algorithm sample.

Agglomerative output	a b c d -
C1 of Thesaurus 1	a b - - -
C2 of Thesaurus 2	a b c - -
C3 of Thesaurus 1	- - c d -
C4 of Thesaurus 3	- - c d e
	3a 3b 2c
	2c 3d 1e

Initially, all words are in the same group and each word has a relation weight (number of lines with occur together) with others words. So, the splits are performed in word relations with score 1 or 0. In the example, letters "a" and "d" occurs just in the first line (agglomerative output line). It probably indicates that there are two distinct concepts that must be separated. Therefore, the algorithm output consists of two groups: {3a, 3b, 2c} and {3c, 3d, 1e}. One may also note that the words in each synset get a weight that is computed from the agglomerative output in the synsets from different thesauri. This weight is then used as a fuzzy membership, as proposed originally by [2]. In this step, the agglomerative original word weight is not used anymore.

One may note that each group of the agglomerative may generate one or more groups. In fact, this step does not produce fewer groups than the agglomerative. In this point, we achieved our first wish, forming large groups and, second, splitting the groups that include two concepts or more. To illustrate this, we return to our real example.

Table 3. Splitting algorithm matrix.

Agglomerative Output	**S1**	**S2**	**S3**	**S4**
amor	x	x	x	
afeto	x	x		
afeição	x			
ardor			x	x
arrebatamento				x
calor			x	x
carinho	x			
fervor				x
paixão		x	x	x
simpatia	x			

In this way, the splitting algorithm output will be a fuzzy synsets as below:

Splitting algorithm output

{amor (2), afeto (2), afeição (1), carinho (1), paixão (1), simpatia (1)}
{paixão (2), ardor (2), calor (2), amor (1), arrebatamento (1), fervor (1)}

5 Results and Discussion

The goal of the methodology presented in section 4 is to group related words in huge synsets to keep the expansive nature of analogies (the more analogies the better for creative writing) and the same time group, through of weights, the words more closely semantics (which facilitates the creative work). Our fuzzy synsets, when represented on a cartesian plane, takes the form of a long tail, ie, a few words have a large weight, whereas many words hold a small weight. Our results show that the words with greater weight are better to lexicalize a concept C than words with less weight. However, the words of lower weight (forming the long tail) serve as a link between different conceptual categories, which shows that the boundaries between very similar concepts have not in fact well-defined contours.

Table 4 shows the noun synsets in the Dicionário Criativo for word *amor*. Each line corresponds to a fuzzy synset and displays only the words with a weight greater than 0.2. Though incorporated in the database, these values are not available through the Dicionário Criativo web interface.

Table 4. Results for word *amor* in the Dicionário Criativo

1. afeição (1), amor (0.95), afeto (0.89), simpatia (0.81), ternura (0.66), apego (0.60), carinho (0.56), inclinação (0.5), benquerença (0.45), benevolência (0.45), querença (0.43), dileção (0.39), amizade (0.39), dedicação (0.33), admiração (0.29), estima (0.29), predileção (0.25), preferência (0.22), derretimento (0.22), meiguice (0.22), idílio (0.22), aferro (0.2), constância (0.2), intimidade (0.2), derriço (0.2), conchego (0.2), estremecimento (0.2), idiopatia (0.2), fraternidade (0.2), [...]
2. paixão (1), ardor (0.684), fervor (0.52), chama (0.44), adoração (0.42), amor (0.39), atração (0.36), devoção (0.34), calor (0.31), êxtase (0.28), enlevo (0.28), flama (0.26), idolatria (0.23), arroubamento (0.21), [...]
3. simpatia (1), afeição (1), afeto (0.84), amor (0.6), amizade (0.56), ternura (0.47), querença (0.47), dileção (0.45), apego (0.45), benquerença (0.45), benevolência (0.43), dedicação (0.39), admiração (0.39), estima (0.37), predileção (0.37), intimidade (0.32), carinho (0.3), preferência (0.3), inclinação (0.26), derretimento (0.21), aferro (0.21), estremecimento (0.21), constância (0.21), idiopatia (0.21), derriço (0.21), idílio (0.21), conchego (0.21), [...]
4. adoração (1), devoção (0.92), veneração (0.64), paixão (0.48), fervor (0.44), idolatria (0.4), amor (0.4), culto (0.36), dedicação (0.36), ardor (0.36), êxtase (0.28), chama (0.28), enlevo (0.28), [...]
5. cuidado (1), amor (0.9), desvelo (0.9), carinho (0.8), dedicação (0.8), zelo (0.7), atenção (0.6), apego (0.4), diligência (0.4), vigilância (0.3), aplicação (0.3), vigília (0.3), afeto (0.3), afeição (0.3), boa vontade (0.2), benevolência (0.2), simpatia (0.2), fraternidade (0.2), ternura (0.2), gosto (0.2), comunhão de sentimentos (0.2), caridade (0.2), inclinação (0.2), adoração (0.2), enfatuação (0.2), devoção (0.2), chamego (0.2), [...]
6. namoro (1), namorico (0.83), flerte (0.8), galanteio (0.63), xodó (0.53), agarramento (0.53), namorilho (0.53), namorisco (0.5), derriço (0.43), namorice (0.36), namoramento (0.36), paleio (0.33), romance (0.33), grude (0.33), chamego (0.33), cera (0.33), pé-dealferes (0.33), prosa (0.33), camote (0.26), suruba (0.26), mormaço 163 (0.26), tribofe (0.26), azeite (0.26), sumbaré (0.26), caso (0.26), namoricho (0.26), amor (0.23), rabicho (0.23), aventura (0.23), amorico (0.2), corte [...]
7. namorado (1), querido (0.81), bem (0.66), amor (0.63), amado (0.59), caro (0.55), predileto (0.55), amante (0.51), derriço (0.48), amigo (0.44), beijoqueiro (0.44), beijocador (0.4), apaixonado (0.37), dileto (0.33), esposo (0.33), estimado (0.29), enamorado (0.29), flerte (0.25), ídolo (0.25), namorido (0.25), frecheiro (0.25), noivo (0.25), zinho (0.25), chichisbéu (0.25), amásio (0.25), cujo (0.25), ficante (0.25), preferido (0.25), favorito (0.22), benzinho (0.22), estremecido (0.22), [...]
8. namorada (1), amada (0.78), querida (0.72), amante (0.51), ídolo (0.48), amor (0.48), predileta (0.45), apaixonada (0.45), deusa (0.42), noiva (0.39), benzinho (0.33), amorzinho (0.33), dulcineia (0.33), anjo (0.3), pequena (0.3), derriço (0.3), bem (0.27), preferida (0.27), amásia (0.27), dileta (0.24), inclinação (0.21), nubente (0.21), pretendida (0.21), concubina (0.21), [...]
9. galanteador (1), namorado (0.9), admirador (0.9), adorador (0.87), galã (0.83), namorador (0.8), amante (0.77), pretendente (0.77), conquistador (0.74), apaixonado (0.71), cortejador (0.67), dom-juan (0.67), namoradeiro (0.58), fã (0.48), adorante (0.41), pretensor (0.41), vegete (0.38), quebra-esquinas (0.38), babão (0.38), bandoleiro (0.38), Casanova (0.38), derriçador (0.38), marrancho (0.38), proco (0.38), jacaré (0.38), amoroso (0.38), noivo (0.25), enamorado (0.25), [...]

10. carícia (1), carinho (1), abraço (0.88), beijo (0.88), galanteio (0.66), galanteria (0.66), galanice (0.67), ternura (0.67), ósculo (0.67), galantaria (0.55), agarramento (0.55), meiguice (0.55), amplexo (0.44), mimo (0.44), afago (0.44), cortejo (0.33), gentileza (0.33), derretimento (0.33), abraçamento (0.33), afeto (0.33), requebro (0.33), agarração (0.33), beijoca (0.33), corte (0.22), volúpia (0.22), agarra (0.22), luxúria (0.22), lisonja (0.22), sexualidade (0.22), voluptuosidade (0.22), orgasmo (0.22), licensiosidade (0.22), gozo (0.22), cópula (0.22), lascívia (0.22), transa (0.22), sexo (0.22), erotismo (0.22), denguice (0.22), amor carnal (0.22), entranha (0.22), donaire (0.22), garbo (0.22), libertinagem (0.22), delicadeza, (0.22), [...]

The expansive feature of our model causes the synsets of our semantic network to be significantly different from synsets of most wordnets. We conducted a simple comparison between our results with those in Onto.PT and OpenWN-PT. Onto.PT is created automatically and, in its construction, it exploited some of the lexical resources we did. Also, compared to other Portuguese wordnets, it is the largest [10]. OpenWN-PT is a more conventional Portuguese wordnet. Table 5 shows the noun synsets in the Onto.PT for word *amor*. Table 6 shows the noun synsets in OpenWN-PT for word *amor*.

Table 5. Results for word *amor* in the Onto.PT

1. aventura, amor
2. amante, amor
3. afecto, amor, paixão, afeição
4. mor, amor, cupido, amança
5. amor, afeição, finura, maciez, dolorimento
6. cuidado, atenção, aplicação, diligência, dedicação, amor, zelo, afeição, desvelo, solicitude, vigilância, carinho, vigília, venida, zêlo, matação, cuidança, cuido, condessilho
7. afeto, afecto, amor, afeição, simpatia, inclinação, fraternidade, admiração, apego
8. culto, afeto, afecto, tenção, amor, homenagem, veneração, altar, devoção, adoração, amorismo, adoramento, latria
9. afecto, amor, estima, afeição, fraternidade, carinho, benevolência, apego, ternura, afecção, afetividade
10. afeto, afecto, amor, estima, amizade, querença, afeição, simpatia, apegamento, atracção, fraternidade, benevolência, apego, veneração, devoção, ternura, afetividade, querência, afeiçoamento, dileção

Table 6. Results for word *amor* in the OpenWN-PT

1. Cupido, Amor	8. querido, amor
2. doçura, querido, amor, anjo	**9.** afeição, amor, apego
3. amor	**10.** afeição, amor

4. amor	**11.** afeição, paixão, amor, Paixão
5. dolorimento, finura, amor, afeição, maciez	**12.** amor
6. amor cortês	**13.** afeição, amor, impressão moral, apego, carinho, afecção, afetividade
7. amante, amor	**14.** montar, pinar, meter, foder, comer, trepar, transar, copular, fazer amor, fazer sexo, dormir com

In general, the three resources exhibit similar numbers of synsets. More precisely, both Onto.PT and Dicionário Criativo have ten synsets for *amor* and OpenWN-PT has fourteen. The differences are in the number of words within each synset and conceptual cut each synset provides.

Notice that the third synset in Onto.PT {afecto, amor, paixão, afeição}, glossed by "good or bad movement of the soul; feeling or emotion of great intensity", and the tenth synset in OpenWN-PT {afeição, amor}, glossed by "a strong positive emotion of regard and affection", are similar to our first synset {amor, afeto, carinho, afeição, simpatia, …}, but our synset has much more words.

Concerning to the conceptual cut each synset provides, notice that the second and the fourth synsets in Onto.PT point to the same concept of personification of love. In OpenWN-PT, the same concept, glossed by "loved one", appears in the second, fourth and seventh synsets, glossed by "any well-liked individual", "a person loved by another person" and "a beloved person", respectively. In Dicionário Criativo this concept is arranged in the same synset 6 {namorado, querido, bem, amor, amado, caro, predileto, amante, derriço, [...]}, again with many more words.

Table 7 shows the numerical analysis of the results.

Table 7. Numerical analysis of the results.

Base	**#Synsets**	**Max(#words in synsets)**	**Mean(#words in synsets)**
Dicionário Criativo	30,605	387	17.91+/-23.81

6 Conclusion

We presented our approach for constructing a complete onomasiological thesaurus taking other thesauri as input. In a near future, we also want to provide fuzzy information inside each synset. This thesaurus is integrated in a database and contributes to the logic and content of the website Dicionário Criativo (www.dicionariocriativo.com.br), which can be seen as a creative writing assistance tool, and currently has more than 250,000 monthly users. The fuzzy onomasiological thesaurus can be accessed through the 'relacionadas' tab, but it is not yet available for download.

Future work may incorporate other dictionaries, thesauri and lexical resources to improve the results. We may also use other heuristics, for instance, to assign different weights to each source of lexical information. It is also in our plans to propose forms of evaluation including questionnaires to be filled by human raters and a deeper comparison with related resources, namely wordnets.

References

1. Azevedo, F.F.d.S.: Dicionário analógico da língua portuguesa: ideias afins/thesaurus. Lexikon (1947)

2. Albuquerque, F.I.: Modelo linguístico-computacional para um dicionário analógico digital. PhD thesis, Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, Araraquara, SP–Brasil (11 2013)

3. Briscoe E. J., Boguraev, B.: Computational Lexicography for Natural Language Processing. London: Longman, (1989)

4. Dias-da-Silva, B.C.: Wordnet.Br: An exercise of human language technology research. In: Proceedings of 3rd international wordnet conference (gwc), 301–303 (2006)

5. Dias da Silva, B.C.; Moraes, H.R.; Oliveira, M.F.; Hasegawa, R.; Amorim, D.A.; Paschoalino, C.; Nascimento, A.C.: Construção de um thesaurus eletrônico para o português do Brasil. Processamento Computacional do Português Escrito e Falado (PROPOR), Vol. 4, pp. 1-10 (2000)

6. Fellbaum, C.: WordNet An Electronic Lexical Database. MIT Press (1998)

7. Florenzano, E.: Dicionário de Ideias Semelhantes. Ediouro (1963)

8. Gonçalo Oliveira, H., Gomes, P.: ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. In Language Resources and Evaluation 48(2):373-393. Springer (2014)

9. Gonçalo Oliveira, H., Gomes, P.: Towards the automatic enrichment of a thesaurus with information in dictionaries. Expert Systems: The Journal of Knowledge Engineering (KDBI special issue), 30(4), pp. 320–332. (2013)

10. Gonçalo Oliveira, H., Paiva, V., Freitas, C., Rademaker, A., Real, L., Simões, S.: As wordnets do português. In: Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) Linguística, Informática e Tradução: Mundos que se Cruzam, Oslo Studies in Language 7(1), 397–424 (2015)

11. Kennedy, A., Szpakowicz, S.: Evaluating Roget's Thesauri. Paragraph 10244 6443 (2008)

12. Marrafa, P.: Wordnet do português: uma base de dados de conhecimento linguístico. Instituto Camões (2001)

13. Marrafa, P.: Portuguese WordNet: general architecture and internal semantic relations. DELTA 18. 131–146 (2002)

14. Maziero, E.G., Pardo, T.A., Di Felippo, A., Dias-da Silva, B.C.: A base de dados lexical ea interface web do TeP 2.0: thesaurus eletrônico para o português do Brasil. In: Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web, ACM (2008) 390–392

15. Naber, D.: OpenThesaurus: Building a thesaurus with a web community. `https://www.openthesaurus.de/download/openthesaurus.pdf` (2004).

16. Paiva, Valeria, Alexandre Rademaker & Gerard de Melo. OpenWordNetPT: An open Brazilian wordnet for reasoning. Em Proceedings of 24th International Conference on Computational Linguistics, 353–360 (2012)

17. Pessek, K.: Dicionário de Palavras Interligadas: analógico e ideias afins. Thesaurus (2010)

18. Roget, P.M.: Roget's Thesaurus of English Words and Phrases... TY Crowell Company (1911)

19. Spitzer, C., Santini, L.: Dicionário analógico da língua portuguêsa: tesouro de vocábulos e frases da língua portuguêsa. Editôra Globo (1953).

20. Wittgenstein, L.: Philosophical Investigations. Oxford: Basil Blachwell (1953)

Universal POS tagging for Portuguese: Issues and Opportunities

Valeria de Paiva and Livy Real

[1] Nuance Communications, USA
[2] IBM Research, Brazil
valeria.depaiva@nuance.co livym@br.ibm.comm

Abstract. Part-of-Speech (POS) tagging consists of labeling every token of a text with its correct morphosyntactic category and is considered by many a solved task in NLP. However, there are many tag systems in use, tags are not very easy to compare, there is no official golden standard and hence comparing performance of different systems is a nightmare, even for English. Much more so for less resourced languages. Recently a collective of researchers decided to tackle this issue and there is a new initiative, the Universal Dependencies project, that is developing cross-linguistically consistent treebanks and annotations for many languages. We look at how the coarse categories of POS tags defined by the Universal Dependencies project would work for Portuguese and describe the issues of aligning them with the POS tags produced by FreeLing, the open source NLP system we use.

1 Introduction

Part-of-Speech (POS) tagging consists of labeling every token of a text with its correct morpho-syntactic category and is considered by many a solved task in NLP, for English, at least. Supervised POS tagging accuracies for English, measured on the Wall Street Journal portion of the PennTreebank, have converged to an impressive 97% [15]. But for languages other than English the situation is not so rosy. For a start, for most languages there are not as many open source POS tagging systems as there are for English. And actually, even for English, the situation is not as good as this simple number might indicate (see [8]).

Nevertheless, work on supervised and unsupervised multilingual tagging is progressing and there is a new initiative, the project Universal Dependencies[3] (UD), that is developing cross-linguistically consistent treebanks and annotations for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning and multilingual parsing research. They aim to produce truly universal POS tags, based on the idea that there is a set of (coarse) syntactic POS categories that work in similar fashion across many, perhaps all, languages. The project is ongoing, having had its first official release (with ten languages) in January 2015. Version 1.1 with eight additional languages was released in May 2015 and subsequent releases are expected every six months, with

[3] http://universaldependencies.github.io/docs/

the next one schedule for May 2016. The guidelines for the UD were released October 1, 2014 and were kept stable for a year. It is expected that guidelines, tags and features may be revised as the discussions unfold and the empirical basis for generalization increases. A ‘laundry list’ of 17 issues (similar to the ones discussed here) was discussed at the Uppsala meeting, as part of Depling 2015, and can be found in `http://universaldependencies.org/issues.html`. It is worth noticing that many remain open questions, as of this writing.

A basic assumption of the Universal Dependencies project, in the words of Nivre [10] is

> [...]that dependency relations hold primarily between content words, while function words are pushed to the bottom of the trees and attached in a flat structure to the content word with which they are most closely associated. This principle is enforced to maximize parallelism across languages, since content words and their relations are more likely to be similar across languages, while function words in one language often correspond to morphological inflection (or nothing at all) in other languages.

While the general principle seems sound and very useful, there are too many details that are not clear cut and seem to deserve a more detailed discussion, in the specific settings of different languages. In this note we look at how these coarse categories of POS tags would work for Portuguese and describe the issues of aligning them with the POS tags produced by FreeLing [12], the open source NLP system which we have been using so far. We are not interested in the POS tagging task in NLP per se, but on whether the tag system proposed by Universal Dependencies project is adequate for Portuguese and if not, how to make it so.

We are also interested in the converse task, the use of pos-tagging to improve lexical resources such as the OpenWordNet-PT [4]. Thus we investigate the state of the existing tags, and then discuss possibilities of implementing new coarser tags similar to the ones in the Universal Dependencies project.

2 Google and Universal tags

To facilitate research in unsupervised induction of syntactic structure and to help standardize best-practices, Petrov, Das and MacDonald [13] proposed a tagset that consists of 12 universal POS categories. As they explain, their reasons were pragmatic: there might be some controversy about what the exact tagset should be, but these categories cover the most frequent parts of speech that seem to exist in most languages. They also developed a mapping from finer grained POS tags for 25 different treebanks to this universal set, showing some level of universality of their tagset. They made the tagset plus mappings[4] available in 2012.

Their universal tagset grew out of the cross-linguistic error analysis based on the CoNLL-X shared task data by [9]. It was initially used for unsupervised part-of-speech tagging by [3] and has since been adopted as a widely used standard

[4] `https://code.google.com/p/uni-dep-tb/`

for mapping diverse tagsets to a common standard, as explained in the Universal Dependencies website.

After extensive discussion, the original set of Google tags was improved to make some distinctions that were missing in the original proposal, but were perceived to be of importance by many. The universal part-of-speech tags (UPOS) are based on the Google universal tagset, which has been extended and redefined from the original 12 to the current 17 tags. The additional 5 tags added are: auxiliary verb (AUX), interjection (INTJ), proper noun (PROPN), subordinating conjunction (SCONJ), and symbol (SYM). In addition, UD also defines a set of 17 universal features that can be used to describe lexical and inflectional properties of words. These features are especially useful for morphologically rich languages. The core feature set is based on Interset [16], an interlingua for morphosyntactic tagsets. It is likely that new features or new feature values will be identified as new languages are added; therefore, the UD format allows additional language-specific features. The full set of 17 tags is listed in Table 1.

open class words	closed class words	other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Table 1. Universal Dependencies tag set

It is worth noticing that Princeton WordNet (PWN) does not list interjections in their open class words. Also it does not deal with any of the closed class words. Since we are mostly interested in the open class words that PWN has content on, we we restrict ourselves to nouns, verbs, adjectives and adverbs in some of our discussion.

Given the weight of the proponents of this suggested lingua franca, Google Research and the Stanford NLP group, it seems very likely that these will become the *de facto* standard in the description and annotation of corpora and hence it makes sense to see how difficult it would be to construct mappings to this standard set from other tagsets. This is a necessary step before defining the universal dependencies for Portuguese, which we also would like to do soon.

3 POS-Tagging Portuguese

There is a considerable amount of work in pos-tagging in Portuguese. In particular recently Garcia and Gamallo have worked exactly on pos-tagging in Portuguese using FreeLing [6,7]. Garcia et al. experiments show that consistency between the training corpus and the dictionary used has a major effect in the

POS tagger performance, at least for the taggers they used. Given the variation between European Portuguese and Brazilian Portuguese, this consistency can be somewhat problematic to attain. Garcia and Gamallo [6] used Brants' tagger to adapt FreeLing for European Portuguese (and for Galician), achieving precision results of up to 96.3%. For Brazilian Portuguese, the work in [1] compared several POS taggers, with best results of 90.25% using the MXPOST algorithm of Ratnaparkhi. Further development, with simplified tagsets, improved the precision up to 97%, but the authors warn that this figure should be taken with some care. As they say, it must be remembered that the corpus used during the training is small, and it is not a representative model of the Portuguese language in general.

FreeLing has a careful discussion of the code it uses for POS tagging on its online documentation. It says it has two different modules to perform POS tagging: a developer needs to decide which method is to be used for a specific application and instantiate the right class. The first POS tagger is the hmm_tagger class, which is a classical trigam Markovian tagger, following the work of Brants [2]. The second module, named relax_tagger, is a hybrid system capable of integrating statistical and hand-coded knowledge, following [11]. The hmm_tagger module is somewhat faster than relax_tagger, but the later allows you to add manual constraints to the model. The manual describes its tagsets for Portuguese in `https://talp-upc.gitbooks.io/FreeLing-user-manual/content/tagsets/tagset-pt.html`. We repeat the 12 tags in Table 2.

open class words	closed class words	other
adjective	adposition	punctuation
adverb		
interjection	conjunction	
noun	determiner	
	number	
verb		
	pronoun	
	date	

Table 2. FreeLing tag set

It is easy to see that FreeLing tagset misses the Universal tags SCONJ (subordinating conjunction), AUX (auxiliary verbs), PROPN (proper nouns), PART (participles), SYM (symbols) and X. In compensation FreeLing has an extra tag for dates, as FreeLing offers a Date Detection Module that already tags time expressions.

We have tried a simple experiment checking a small collection of sentences of the *Floresta Sintáctica* corpus [5], to see which issues we would need to deal with, both with old and new tags. We discuss some of these issues below, treating them as questions, as we have not decided on how to proceed yet.

4 Issues with POS tagging

We have taken a small sample of sentences in Portuguese, from Brazilian newspaper articles and analysed them. We want to use the lexical resource OpenWordNet-PT as a lexicon for further processing, thus we check which words FreeLing tags as nouns, verbs, adjectives and adverbs in these sentences and we try to find them in OpenWordNet-PT. We check which ones of these words are present in the OpenWordNet-PT, with the right part-of-speech and the right meaning and which ones are not and why not.

Some researchers will say that "POS tagging is a mostly (if not purely) syntactic task". We disagree, the task is syntactic for sure, but it has a huge component of semantic information involved. As Zeman [16] explains most of the time a tag is a "compressed representation of a feature-value structure", hence the use of the term "morphological tag" for them. The goal of POS tagging task for us is to make sure that the expected semantics of the sentences is respected by the segmentation/tagging interplay.

The idea in this note is both to improve the lexical resource, by checking that it has the required words with appropriate parts of speech and meanings, but also to verify the quality of the POS tagging code, by checking how many correct tags it gets, for each sentence. Thirdly and most importantly, we want to check the adequacy of the proposed Google tags for Portuguese. This implies reviewing and discussing the relevant issues that are still undecided on that project. Some issues are practically very important, even if theoretically not so. For instance, it is recognized that having the full sentence without annotations as part of the treebank is very useful: for machine learning and linguists. Standardizing on having such and with a single, uniform label is easy, needs to be done, but does not reflect any theoretical insights. However many of the issues under discussion do reflect theoretical differences (e.g. how to annotate light verb constructions, how to annotate pronominal verbs, etc).

For the sentence *Aqui era o quarto, pobre, limpo, simples e acolhedor*[5] we would like FreeLing to detect that *quarto* ('room') is a noun, that *era* ('be') is the verb, that *aqui* ('here') is an adverb and that *pobre, limpo, simples, acolhedor* are adjectives. FreeLing recognizes *quarto* as the adjective 'fourth', not as a noun[6], but the other content words are properly tagged. All the content words are in the lexicon with the appropriate parts of speech.

For the sentence *Os jogadores se dividem pelos dez quartos do alojamento, equipados com frigobar, ar condicionado, televisão e telefone*[7] we would like FreeLing to detect that *jogador, quarto, alojamento, frigobar, televisão e telefone* ('player', 'room', 'lodge', 'minibar', 'television', 'telephone') are nouns, that *dividir, equipar* ('share', 'equip') are verbs and that *dez* ('ten') is a numeral. But

[5] 'Here was the room, poor, clean, simple and cozy.'

[6] FreeLing does have *quarto* as a noun in its dictionary, it just prefers the adjective part of speech in this example.

[7] 'The players are sharing ten rooms in the lodge, equipped with minibar, air conditioning, television and telephone.'

the POS tagging only recognizes *dez quartos* as a unit in this example. We would also want the tagging to see *ar condicionado* as a multi-word expression (mwe). If the tokenization is wrong and *ar condicionado* comes as two tokens, how do we measure the error? Is it one error or two? Lastly, we could want the tagger to know the determiners and the prepositions in the sentence, but for the purpose of the exercise in this note and for checking the lexical resource, we only need to check the open class words of nouns, verbs, adjectives and adverbs. So we restrict our attention to these.

Several questions present themselves, when we start to look at this set of sentences. Some of these questions are language specific, but mostly they are about the POS tagging state of art and how to define it, so that it is parallel in many languages.

What should we do with out of vocabulary (OOV) words? Which is the most perspicuous tag for them? They can be of several kinds, for instance colloquialisms (*cê tá indo aonde?*/ 'u going where?'), foreign words used in their original language (*teens, blues*), regionalisms (*piracema*/'a natural phenomenon when fish swim up river'), neologisms (*frigobar*, 'a hotel small refrigerator'; *tuitar* 'to tweet'), acronyms (IBM, FSE, OIC), etc. Most dictionaries would not have these words, but they do show up in corpora and we need to decide how to deal with them. Taggers usually have defaults and one needs to check that they are appropriate. Tagging *tá* (the verb *estar* can be used for *Yes!*) as an interjection is very reasonable, but not always. FreeLing's 'Unknown Word Guesser Module' seems to do a good job most of the time.

More importantly, there is also the out-of-vocabulary issue that is truly a failing of the lexical resources and these should be counted separately, perhaps. A word might be missing from the processing dictionary (and be treated as a unknown word) and/or can be known by the processing, but be missing the semantic meaning in the OpenWordNet-PT. In the previous example the word *frigobar* (for a refrigerator in a hotel room) was missing both from the FreeLing dictionary (it was guessed as a verb), and from OWN-PT. The word *vão* ('hole', 'opening') was missing in the OWN-PT, as a noun, in the sentence *Para melhorar a ventilação, podem ser criadas janelas nos telhados ou pequenos vãos.com telas para evitar a entrada de insetos*[8] but it also did not show up in the Freeling processing, due to a tokenization error.

What should we do with Named Entities? Should they be tagged as proper nouns or nouns? The universal tags have proper nouns, and FreeLing does have the subcategory, so making the change is not difficult. Some named entities are present in our lexicon at the moment, e.g. *Charles de Gaulle*, many will not be, e.g. *Barak Obama*. Some might be abbreviations, such as IBM and NY; some might look like multiword expressions, like *Ministério da Fazenda* (Department of Finance). Some abbreviations are fairly well-known, such as ONU (Organização das Nações Unidas or UN, for United Nations), and OMS (Organização Mundia da Saúde or 'WHO', World Health Organization). Others,

[8] 'To improve ventilation, windows or small openings can be created on rooftops, with screens to prevent the entry of insects.'

like *FSE*[9] in the sentence *Na época, o então ministro da Fazenda, Fernando Henrique Cardoso, fez um pronunciamento em cadeia nacional para anunciar a intenção do governo de destinar o FSE a investimentos sociais*[10], are not so well known.

Recognizing named entities is, of course, a problem on its own, but they have to be classified as well. Which types of named entities should we have as a bare minimum? Most systems have types for *person, location, organization* and a category *other* seems sensible. But there is also the discussion of which of these named entities should you have in your lexical resource. Since our lexical database OpenWordnet-PT comes from Princeton's Wordnet, only a few named entities are available in that resource. We need to address the issue of how to deal with named entities, since this kind of information could also be extracted from an encyclopedic resource, such as Wikipedia, DBpedia or GeoNames, as discussed in [14].

Which kinds of numbers in the same tag? Most of the tag systems have numerals, like the *dez* ('ten') in *dez quartos* ('ten rooms') in the sentence above. But which other kinds of mathematical entities should be in the same tag? FreeLing has a special tag *date* which is not in the UD tagset. A recent discussion in the issues tracker for the Open Dependencies project showed that Germanic languages differ from Romance languages as to how they refer to dates, for instance. The discussion and (preliminary) conclusions are recorded at `https://github.com/UniversalDependencies/docs/issues/210`. A similar, but not finalized discussion, is going on about hours: should *20:30* in *he met me at 20:30* be tagged as a noun or as a numeral? What if you write it as *20h30*? Does it matter if you say the 'hours/horas' or not when you read the sentence? Similarly the UD tagset has the tag SYM (symbol) to be used for percent signs and other mathematical symbols, but FreeLing has not.

What to do with what are clearly typos in the text? For instance the full period in the example *Para melhorar a ventilação, podem ser criadas janelas nos telhados ou pequenos vãos.com telas para evitar a entrada de insetos*[11] that should perhaps be a comma. For the Portuguese corpus *Floresta Sintáctica* there were guidelines that enforced the non-modification of the sentences in the corpus. Corpora in general will have typos and mistakes and normally this is not an issue. But when the corpus is supposed to be used as the golden standard from where all the community will learn its annotations, it can be frustrating. Especially when it has many words that do not exist in the original language, that are simply misspellings of true words.

[9] *Fundo Social Europeu*, 'European Social Fund' (ESF).

[10] At that time, the then finance minister, Fernando Henrique Cardoso, made a statement on national television to announce the government's intention of allocating the EFS to social investments.'

[11] 'To improve ventilation, windows or small openings can be created on rooftops, with screens to prevent the entry of insects.'

What to do with MWEs? How to deal with them minimally? As the Universal Dependencies site explains, when discussing tokenization[12] "in principle, the lexicalist view could also be taken to imply that certain multiword annotations should be treated as single words in the annotation. So far, however, multiword expressions are annotated as such using special dependency relations, rather than by collapsing multiple tokens into one." While following their lead is the easiest option, given this work's origin in using OpenWordnet-PT and Princeton's WordNet, many MWEs are already lexicalized, like "air conditioning", for example. Not using such MWEs seems a step backwards, semantically. Particularly when it comes to adverbial expressions, not to treat them as MWE seems a seriously bad idea. Do we need to be able to separate noun-noun compounds, like *assessor de imprensa* at this level or not?

but experience shows that coarser tags get better numbers.

What to do about reported speech and quotations? Many other grammatical issues are still being discussed. As far as verbs are concerned, the working group decided that marking auxiliar verbs as distinct from main lexical verbs was important. But many questions remain: how to mark light verbs? What is the extent of the auxiliary verbs?

To start to survey these issues and determine reasonable ways of measuring precision and recall for POS tagging, a small corpus of twenty five short sentences was extracted from the manually corrected portion of the Bosque corpus and analyzed. The main conclusion, so far, is that the questions discussed above need addressing and that more experimentation with adapters for FreeLing is necessary.

5 Experiment and numbers

So far we have performed a very small experiment, devising our own golden standard, where we disagreed with the Bosque tags, whose numbers can be summarized thus:

	FreeLing	Bosque	Golden
sentences	25	25	25
tokens	720	716	714
nouns	131	151	142
verbs	84	82	82
adjectives	36	43	42
adverbs	18	20	20
proper nouns	57	43	46
numbers	22	21	20
dates	7	0	0
symbol	0	0	3

Table 3. Comparing tags

FreeLing does not have the 5 new tags added to the Google Tags by the Universal Dependencies project. We would like to have them. FreeLing has one

[12] `http://universaldependencies.org/u/overview/tokenization.html`

tag that both Bosque and the UDs do not consider, a special tag for dates, which we think is not necessary as a morphological tag.

The fact that dates are separate in FreeLing explains some of the differences in number of nouns as in, e.g. the date *31 de janeiro*, *janeiro* is a noun. Temporal expressions are also treated differently and are a topic under discussion in the Universal Dependencies forum.

A well known issue occurs with participles: sometimes they are tagged as verbs, sometimes as adjectives and the difference is not so easy to detect. World knowledge can play a part even on this shallow level of processing: the sentence *BRASÍLIA Pesquisa Datafolha publicada hoje revela um dado supreendente: recusando uma postura radical, a esmagadora maioria (77%) dos eleitores quer o PT participando do Governo Fernando Henrique Cardoso*[13], FreeLing tags *quer* as a conjunction, when it is clearly a form of the verb *querer* (to want).

Another small difference between tagsets is treating the percent sign `%` as either a noun or as a symbol. We follow the UD tags and think this should be a symbol, just as the dollar sign `$`. Altogether FreeeLing's performance is very good, as we are comparing it to humans and these already have differences amongst themselves.

But most of the disagreement is on how to tokenize multi word expressions (MWEs) and especially entity names, both in UDs and in FreeLing. (There are also many differences on how to tokenize and classify prepositions and determiners, but we are not interested in those, for the time being.) There is one adjective (*italiano*) that the Bosque treats as a noun, maybe simply an oversight. There are two nouns that our golden standard considers proper nouns (*Lua, Terra/Moon, Earth*) while Bosque thinks of them as common nouns.

6 Conclusion

This preliminary note puts forward the idea of adapting FreeLing to use the POS tags of the project Universal Dependencies and discusses some of the issues involved. While it seems clear that POS tagging, named entity recognition and tokenization are inter-related tasks, it is not so clear to us which ways will lead to better performance. The ever present problems of recognizing MWEs, compounds and OOV words, as well as the ambiguity issues are still plaguing us very much, but some progress seems to have been made and more of it can be made, if multilingual corpora, tags, and dependencies can be aligned. As a next step we want to run FreeLing in the whole Bosque corpus and adapt the UD dependencies scripts to check for the inconsistencies between Zeman's conversion of the Bosque dependencies in `https://github.com/UniversalDependencies/UD_Portuguese` and our own results, as well as the official guidelines. Aligning tags and dependencies, with our aim firmly set on semantics, is our goal.

[13] Brasilia Datafolha research published today reveals a surprising fact: refusing a radical posture, the absolute majority of the electors wants the PT participating in the Government of Fernando Henrique Cardoso.

References

1. Rachel V. Xavier Aires, Sandra M. Aluísio, Denise C. S. Kuhn, Marcio L. B. Andreeta, and Osvaldo N. Oliveira. Combining Multiple Classifiers to Improve Part of Speech Tagging: A case study for Brazilian Portuguese. In *Proceedings of the Brazilian AI Symposium (SBIA2000)*, pages 20–22, 2000.
2. Thorsten Brants. TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
3. Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL*, 2011.
4. Valeria de Paiva, Alexandre Rademaker, and Gerard de Melo. OpenWordNet-PT: An open Brazilian wordnet for reasoning. In *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.
5. Cláudia Freitas, Paulo Rocha, and Eckhard Bick. Floresta sintá (c) tica: Bigger, thicker and easier. In *Computational Processing of the Portuguese Language*, pages 216–219. Springer, 2008.
6. Marcos Garcia and Pablo Gamallo. Análise morfossintáctica para português europeu e galego: Problemas, soluçoes e avaliaçao. *Linguamática*, 2(2):59–67, 2010.
7. Marcos Garcia, Pablo Gamallo, Iria Gayo, and Miguel A. Pousada Cruz. Pos-tagging the web in Portuguese. national varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, 53(0):95–101, 2014.
8. Christopher D Manning. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer, 2011.
9. Ryan T McDonald and Joakim Nivre. Characterizing the errors of data-driven dependency parsing models. In *EMNLP-CoNLL*, pages 122–131, 2007.
10. Joakim Nivre. Universal dependencies for swedish. In *In Proceedings of the Fifth Swedish Language Technology Conference (SLTC) Uppsala University, 13-14 November 2014*, 2014.
11. Lluís Padró. A hybrid environment for syntax-semantic tagging. *arXiv preprint cmp-lg/9802002*, 1998.
12. Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May 2012. ELRA.
13. Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.
14. Livy Real, Valeria de Paiva, Fabricio Chalub, and Alexandre Rademaker. Gentle with gentilics. In *Proceedings of the Workshop on LREC*, 2016.
15. Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Procs. of the 2003 North American Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. ACL, 2003.
16. Daniel Zeman. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*, 2008.