Semantic Links for Portuguese

Fabricio Chalub[1], Livy Real[1], Alexandre Rademaker[1,2], Valeria de Paiva[3]
[1] IBM Research Brazil, [2] FGV/EMAp Brazil, [3] Nuance Communications USA
fchalub@br.ibm.com, livym@br.ibm.com, alexrad@br.ibm.com, valeria.depaiva@gmail.com

Abstract

This paper describes work on incorporating Princenton's WordNet morphosemantics links to the fabric of the Portuguese OpenWordNet-PT. Morphosemantic links are relations between verbs and derivationally related nouns that are semantically typed (such as for instance "*tune-tuner*" — in Portuguese "*afinar-afinador*" – linked through an *agent* link). Morphosemantic links have been discussed for Princeton's WordNet for a while, but have not been added to the official database. These links are very useful, they help us to improve our Portuguese WordNet. Thus we discuss the integration of these links in our base and the issues we encountered with the integration.
Keywords: Wordnet, Morphosemantics, Portuguese

1. Introduction

Fellbaum, Osherson and Clark (Fellbaum and Clark, 2007) discuss how, to aid automatic reasoning using WordNet, it would be useful to classify and label the relations between nouns and verbs that are derivationally and semantically related. Many traditional paper dictionaries include morphological derivations between verb and noun pairs, but they tend to simply list them as run-ons without any information on their meaning. For example, the large-scale database of categorial variations of English lexemes, known as CatVar (Habash and Dorr, 2003) contains some 100,000 unique English word forms; however, no information is given on the meanings of these words. Similarly, for Portuguese, the "Portal da Língua Portuguesa"[1] lists more than 5,400 deverbals with their respective verbs, but again has no meanings associated.

By contrast, the work on WordNet itself pays special attention to meanings as logical constructs. They say "We are currently working to transform WordNet into a Knowledge Base that better supports such reasoning and inferencing" (Fellbaum and Clark, 2007, p.02). Thus it is surprising to us that for several years they have provided a collection of files, manually checked, with almost 17k typed links, relating verbs and nouns, but that these links have not been made part of the official distribution of WordNet, nor of the online version, available for users' consultation.

For many NLP tasks and also for people trying to construct semi-automatically wordnets in other languages such a resource would seem very useful. We started our own version of morphosemantic links, not knowing about the Princeton Wordnet ones to begin with. The method we used to produce our original morphosemantic links in Portuguese was somewhat tortuous, as deciding what semantics types are relevant and how to label them are not trivial tasks. When we realized the existence of the Princeton morphosemantics links we decided to integrate our previous work with theirs, as described below.

2. OpenWordNet-PT and NomLex-PT

Our main reason for producing lexical resources is to capture the meaning of natural language expressions in representations suitable for performing inferences, that is, we want to do computational semantics.

Semantic information in the shape of lexical resources is relevant for several NLP tasks, such as machine translation, information extraction and even to assist the creation of annotations for machine learning tools. We have been working on what we consider our first long-term and wide-scale lexical resource OpenWordNet-PT (de Paiva et al., 2012) for some years now. OpenWordnet-PT is a freely available wordnet for Portuguese, browsable at `http://wnpt.brlcloud.com/wn/` and downloadable at `http://github.com/own-pt/`.

We have also been working on a smaller lexical resource, a lexicon of nominalizations in Portuguese called NomLex-PT (de Paiva et al., 2014), embedded into OpenWordnet-PT and freely available for download at `http://github.com/own-pt/nomlex-pt`. NomLex-PT offers a list with some 4,240 pairs of related verb/noun forms in Portuguese. To construct NomLex-PT, we semi-automatically translated the original English NomLex (Macleod et al., 1998), the French Nomage (Balvet et al., 2011), the Spanish AnCora-Nom (Peris and Taulé, 2011) and manually verified the pairs acquired. Then worrying that we would be missing truly Portuguese deverbals, we investigated a collection of Portuguese corpora (the AC/DC corpora (Costa et al., 2009) collection) to complete our collection of nominalizations, obtaining the resource described in (de Paiva et al., 2014).

Nominalizations, nouns formed from other part-of-speech words, such as for example "construction" and "government", constitute one of most well known polysemous and problematic issues of formal theories in Linguistics. These nominals have a clear morphological link with the related verb, but their meanings are not automatically derivable from the meaning of the base verb nor are they directly obtainable from the composition between the meaning of the base verb and its suffix. "Government", for example, is formed by the suffix "-ment" which, in general means "the event of doing X", but "government" (and the Portuguese *governo*) has several possible meanings: the event of governing, the result of governing, the period of time some governing happened, the people that govern, etc.

Nominalizations have been heavily investigated in theoret-

[1] `http://www.portaldalinguaportuguesa.org/`.

ical and computational linguistics. (Chomsky, 1970) was one of the first works which have pointed out the phenomenon and proposed a whole theory based on the lexicalization of deverbals nouns, the Lexicalist Hypothesis. Many linguistic works have paid attention to nominalizations (Clark, 1979; Pustejovsky, 1995; Alexiadou, 2001; Brandtner, 2011; Jezek and Melloni, 2009; Jezek and Melloni, 2011) and their logical polysemy has been a hot topic in lexical semantics, as it poses challenges to formal theories and to computational treatments of deverbals alike. The computational treatments try to automatically predict the meaning of deverbals and to uncover the implicit verbal arguments when confronted with a nominal deverbal (Gurevich et al., 2008; Gurevich and Waterman, 2009).

The work in (Real and Retoré, 2014) argues that one needs to have nominalization meanings encoded in the lexicon, as their formation do not follow a general semantic pattern. The work in (Fellbaum and Clark, 2007, p.04) remarks that the most regular formation of nominalizations in English, the agentive pattern brought about by the suffix "-er" or "-or", works for only two thirds of their listed examples. Thus a computational treatment that hopes to provide semantics for all those nominals must have this information encoded in a lexical resource.

One major issue here is the polysemy of both verbs and nouns, related by morphosemantic links. Table 1 summarizes the number of monosemous (a single sense) nominalizations for verbs and nouns, described in NomLex-PT and in the morpholinks of PWN. That is, from 4,238 nominalizations in NomLex-PT, we have only 315 nominalizations where both the verb and the noun are monosemous, less than 10% of the listed pairs. The numbers for English pairs are even smaller, only 717 of the almost 17 thousand morphosemantic links have unambiguous verb and noun in English. The analysis of the effect of polysemy in both the English and the Portuguese resources is another issue where we believe this work on morphosemantic links will be helpful with in the very near future.

	nomlex (OWN-PT)	morphosemantic links (PWN)
verb	963 (22.7%)	1,208 (7.1%)
noun	1,202 (28.3%)	2,832 (16.6%)
both	315 (7.4%)	717 (4.21%)
total	4,238 (100%)	16,995 (100%)

Table 1: Monosemy of verbs and deverbal nouns in Portuguese and English

The highly polysemic nature of nominalizations has being heavily investigated by formal linguistics, specially by lexical semanticists, as (Asher, 2011), that are interested on the behavior of nominalizations, mainly in copredication contexts — that are contexts in which one single token brings about more than one semantic pattern to the sentence. An example of copredication is "The heavy translation was revised twice", where "translation" concomitantly means the physical result of the act of translating (characterized by the adjective "heavy") and its informational content (that is what can be revised) . There are many works on this subject and, for now, there is no uniform treatment to understand, and more than that, to predict, their polysemy.

The polysemic pattern "event/result" — present in "construction", "development" and "accomplishment" — is surely the most recurrent in, at least, West Germanic and Romance languages, as English, Dutch, Portuguese, Spanish, French, etc. But there is no work that correctly predicts when a nominalization would have only one of those two possible meanings or both (Real and Retoré, 2014). Since the polysemy of nominalization is still an unsolved challenge even for theoretical analysis, we believe that lexical resources as wordnets should encode systematically this information, however considering the high degree of polysemy we did not expect that it would be an easy task.

One established way of codifying these kinds of semantic information between nominalizations and their corresponding verbs in a lexical resource is using the idea of morphosemantics links. These relate a noun and a verb senses, adding a label that indicates the kind of relationship the lexical items have between themselves. Princeton's Wordnet team did one project to obtain such links in 2007, but they do not offer those semantic links in their main database. The list is available as a standoff file with 16,995 typed links between noun/verb senses. [2] The semantic relations used in this list are: agent, body-part, by-means-of, destination, event, instrument, location, material, property, result, state, undergoer, uses and vehicle; which is a slightly different list from the one discussed in the paper (Fellbaum and Clark, 2007).

Here we describe how we added to OpenWordnet-PT a set of morphosemantic links between noun/verb senses pairs. As basic resources we have used the previous work done by the Princeton's Wordnet team, described in (Fellbaum and Clark, 2007), and our own NomLex-PT. Rather than simply attempting to manually project the verb-noun pairs between words forms from Nomlex-PT to links between particular senses of those words, a difficult task given the fine-grained nature of Princeton's synsets and the state (not fully curated) of our Portuguese OpenWordnet-PT, we decided to use Princeton's morphosemantic links as helping data to both discover issues with OpenWordNet-PT synsets and to help to project the links from NomLex-PT. Thus the work we describe here consists in adding to the pairs of translated morphosemantic links, that is to pairs of senses of verbs/nouns in Portuguese, a label from Princeton's table and making such a triple, a link of the OpenWordnet-PT.

3. Semantic Links

Before describing the several different uses we see for the newly properly attached collection of morphosemantic links to our Portuguese wordnet, we thought it might be relevant to discuss some examples of these linkages. The paradigmatic examples given by the Princeton's team are a good place to start discussing the morphosemantic links and how they can be incorporated into a Portuguese wordnet. They exemplify

[2] Actually the WordNet Princetons file has 17,739 links, but some of them are repeated. This list can be downloaded from `http://wordnet.princeton.edu/wordnet/download/standoff/`.

their relations via the Table 2, as expounded in their `morphosemantic-links-README.txt` file.

Relation	Example
agent	*employ-employer*
body-part	*abduct-abductor*
by-means-of	*dilate-dilator*
destination	*tee-tee*
event	*employ-employment*
instrument	*poke-poker*
location	*bath-bath*
material	*insulate-insulator*
property	*cool-cool*
result	*liquefy-liquid*
state	*transcend-transcendence*
undergoer	*employee-employ*
uses	*harness-harness*
vehicle	*kayak-kayak*

Table 2: Morphosemantic Relations from PWN

But out of these 14 examples only three would work easily in Portuguese. The pair *employ-employer* (*empregar-empregador*) also has the right relationship *agent* in Portuguese. We also have the *undergoer* relationship in this case *employ-employee* (*empregar-empregado*) and the *result* link for *liquefy-liquid* (*liquidificar-líquido*). But all the others do not seem to work.

Either there is no direct translation (we do not have a specific verb for "teeing" in Portuguese) or the verb is rarely, if ever, used, *kayak-kayak* (*caiacar-caiaque*). More commonly, as in the example "bath-bath", the relationship is ambiguous, as it happens in English too. "Bath" can be thought of the place/location where one bathes (*banheira* in Portuguese) but can be the artifact or instrument used to bathe, like in "baby bath". Of course "bath" can also be the result/event/action of bathing (*banho* in Portuguese). We also use in Brazil *banheiro*, for bathroom, which is *casa de banhos* in European Portuguese.

While the pair "*abduct-abductor*" can be used in the muscular sense in Portuguese — {01449427-v abduct, *abduzir* — pull away from the body; "this muscle abducts"}, it seems more commonly used in its kidnapping sense, {01471043-v snatch, kidnap, abduct, nobble, *abduzir, sequestrar, raptar* — take away to an undisclosed location against their will and usually in order to extract a ransom; "The industrialist's son was kidnapped"}, where the role would be *agent*.

However looking up the pair "*dilate-dilator*", we see one of the first applications of the work morphosemantic links do for us. They help us to complete empty synsets in Portuguese, especially for verbs that are not very commonly used and hence were not picked up by the Wikipedia-based construction of the OpenWordNet-PT. Thus looking up "dilate" we find {00305537-v dilate, distend *expandir, dilatar* — become wider; "His pupils were dilated"} and realize that the verb *distender* is not in the OpenWordNet-PT. Then looking up the empty (in Portuguese) synsets 00257087-v and 00256862-v, we realize that we should also complete them with the verb *distender*:

1. {00257087-v distend — cause to expand as it by internal pressure; "The gas distended the animal's body"}
2. {00256862-v distend — swell from or as if from internal pressure; "The distended bellies of the starving cows"}
3. {00305537-v dilate, distend *expandir, dilatar* — become wider; "His pupils were dilated"}

Moreover, when searching for "dilate" in English, we see that we have the synset {00955601-v expand, expound, dilate, expatiate, lucubrate, flesh out, elaborate, enlarge, exposit elaborar}. This synset fails our guidelines, which try to have a similar number of words in English and in Portuguese. This synset has 9 words in English and a single one in Portuguese. So the Portuguese synset should be completed with verbs corresponding to these other elements: we can add *elocubrar, expor, expandir* to this synset and while *expor, expandir* were already in the lexicon (in other synsets), we were missing *elocubrar* ("locubrate") altogether. This application of the morphosemantic links, helping to complete wordnets in other languages, was the one original application of this work, described for Turkish in (Bilgin et al., 2004) and for Bulgarian in (Stoyanova et al., 2013).

4. Opportunities and Challenges

From almost 17,000 morphosemantic links available from Princeton's wordnet, we could automatically derive around 2,700 morphosemantic links for OpenWordnet-PT. A first checking of these links pointed out to an unbelievably high accuracy: looking at the words of the extracted links in Portuguese and the labels (associated to approximately half of these links), we had around 96% of correct morphosemantic links in Portuguese. Later on a more thorough investigation of the synsets themselves revealed many issues.

Simply reading words in synsets, we could superficially check the correctness of links and these look very good. However, simply reading the glosses (and assuming glosses correct), we could and did miss more subtle errors. For example we have the link *agent* in *escrever-escritor* between the verbs in {01744611-v publish, write — have (one's written work) issued for publication; "How many books did Georges Simenon write?"} and the noun corresponding to the synset {10794014-n author, writer — writes (books or stories or articles or the like) professionally (for pay)} which looks correct, as a writer (*escritor*) is an agent who writes (*escreve*) for payment. However this verb synset turns out to be about publishing books, not about writing them. And the nominalization *publish-publisher* does not seem to exist in Portuguese.

We then manually checked 10% of the proposed relationships, comparing the synsets that the relations link, the type of the label inherited from the PWN work and all the other possible synsets in PWN that contain the related words. From 261 manually checked pairs, we found 60 wrong pairs. (We intend to continue checking the other links extracted, but believe that this error rate is reasonable.)

The quality of these links is still surprisingly high, especially considering the discussion above that indicates that

the links could fail for many different reasons: our Portuguese wordnet is still missing many translations for lexical items, there are verbs that only make sense in English, there are verbs that are only verbs in Portuguese (e.g. to give or receive solidarity to/from someone (*solidarizar*)), PWN relations are not so well curated, we do not expect always to have the same nominalizations in English and in Portuguese, etc.

One example of when things work well is, for instance, the link labelled *event* between the verb Americanize (*americanizar*) (synsets 00409643-v and 00410406-v) and the nominal Americanization (*americanização*) (synset 13429888-n). One may perhaps dispute the need for two verbal synsets for the verb *Americanize* (one for becoming American yourself, the other one for making something or someone more American) in PWN but it is clear that the meanings correspond exactly to the ones in Portuguese in the nomlex-pt link *americanizar-americanização*. Note that in this case both the verbal synsets and the nominal synset have a single word form, if one does not consider the merely orthographic difference between "Americanize" and "Americanise".

Other cases where things seem to work well are the ones where the senses are narrowly defined, e.g. allocate (*alocar*) (synset 02234087-v). If the morphologically related nouns associated with a pair of directly translated verbs exist in both languages, they seem to have a similar relation, in this case to the nominal allocation (*alocação*) (synset 13289467-n). This is a disputable case for the label though, PWN has *undergoer* but *event* or *result* might be better. Differently from the last example, the synset allocate (*alocar*) (synset 02234087-v) has not just a single word, it has two: "allocate" and "apportion", and both of those words are morphologically related to nominalizations, "allocation" and "apportionment". Since we have no Portuguese pair that directly corresponds to "apportion" and "apportionment", morphosemantic links generated from this pair, if any, are considered as connecting *alocar-alocação* in Portuguese. Although PWN keeps the link between {"apportion","apportionment"} and {"allocate","allocation"}, the word form "apportioning" in the {01083645-n allocation, allotment, apportionment, apportioning, parceling, parcelling, assignation — the act of distributing by allotting or apportioning} does not have a morphosemantic link relating it to the verbal synset, nor a morphological relation. This shows us that even between the PWN synsets, the morphosemantic relations are not very consistent.

There are however, many kinds of bad links and we are, of course, more interested on the failures than on the successes of our heuristics. There are bad links where the label seems wrong, but the synsets are connected. For example we have an *agent* link between *suspect-suspect* (*suspeitar-suspeito*) but the noun {09762101-n suspect, defendant *acusado, réu, suspeito, argüido* — a person or institution against whom an action is brought in a court of law; the person being sued or accused} is not the agent, but some kind of patient or undergoer of the verb suspect {00924873-v suspect *suspeitar* — hold in suspicion; believe to be guilty}.

Another example of a wrong link is the one between *sorrir-sorriso* in Portuguese, created via the pair "*grin-smile*" in English. Clearly the pair *sorrir-sorriso* is a correct nominalization in Portuguese, which corresponds to the pair "*smile-smile*" in English. Unfortunately the system gets the pair *sorrir-sorriso* also from a relationship between the pair "*grin-smile*", as the verb {00029025-v grin — to draw back the lips and reveal the teeth, in a smile, grimace, or snarl} is not necessarily related by an event to the noun "smile" (in the sense of *a facial expression characterized by turning up the corners of the mouth; usually shows pleasure or amusement*), but they get connected by the rule 1 (see next section). However, there is a not so subtle difference in meaning in English. The verb "grin", which can be "sorrir", is also used for making a more general facial expression, which we might call a "careta" in Portuguese. Only a pleasant facial expression corresponds to a "sorriso" in Portuguese. Portuguese does not have the verb for the general facial expression associated to a "grin", the closest in Portuguese would be the verbal expression make grimaces (*fazer careta*). This false positive or wrong candidate came about as the word "grin" appears both in the verbal and in the nominal synsets for smiling in PWN {06878071-n grin, grinning, smile, smiling — a facial expression characterized by turning up the corners of the mouth; usually shows pleasure or amusement}. Both pairs of words (*grin-grin*, *smile-smile*) are related through derivationally related links, which indicate the morphological relation between them. However, only one of these links the pair (*smile-smile*) should create a morphosemantic link in Portuguese.

We found several bad links following this pattern, in which there are obvious morphological relations between the words, but only some of the links in PWN correspond to a morphosemantic link in Portuguese.

There seems to be several reasons for the existence of 'wrong' candidate pairs in Portuguese. To begin with, morphosemantic links are between morphologically related words, but the converse is not true: not all morphologically related forms are also semantically related. An easy example in Portuguese is the pair *procurar-procuração*, where the verb means "to seek", while the noun means "powers of attorney". The nominalization is lexicalized and it has lost any connection with the verb, but they are clearly morphologically related.

Then English has many near-synonyms that are translated to a single word in Portuguese. An example here is the verbal synset {01684337-v sculpt, sculpture — esculpir}. This pair of verbs will cause multiplication of links, but no semantic issues. More importantly, our initial heuristics, described in detail in the next section, have some drawbacks. When PWN offers a large number of synsets containing the same words and there are many morphosemantic links involved, our heuristics do not always link the right word forms within synsets. Thus we obtain the link *purificar-purificação* from the connection between the synsets {00475183-v purify, sublimate, make pure, distill — remove impurities from, increase the concentration of, and separate through the process of distillation; "purify the water"} and {13468306-n distillment, distillation — the process of purifying a liquid by boiling it and condensing its vapors}. Clearly the pairs *purify-purification*, *sublimate-*

sublimation, *distill-distilation* should all be amongst the sources of the nominalization pairs in Portuguese. The pairs *purificar-purificação*, *sublimar-sublimação* and *destilar-destilação* are clearly nominalizations in Portuguese. But we should not obtain *purificar-purificação* from the synset that has only *distillment-distillation* as a nominal.

These examples show that the kind of manual checking we have done so far is not enough to guarantee correctness of all the links. We can have the same word pair in Portuguese *experimentar-experiência* (translated as *experiment-experience*) and the same kind of semantic link *event* with different senses in English: the synset 02532886-v is the verb *experiment* in the sense of the act of "trying out" a new experience, while 01771535-v is about the feeling one experiments and both will have the the same word *experiência* as their nominalization, but with very different meanings.

An interesting example, that shows a different use of the work described here, is improving PWN itself. When we look at the morphosemantic link labelled *event* between the Portuguese pair *passear-passeio*, that means take a walk (*a walk*), we found from one single NomLex relation (*passear-passeio*) two links candidates: the pairs *amble-amble* and the *stroll-stroll*, that are in different verbal synsets but very closely related: {01917980-v stroll, saunter — walk leisurely and with no apparent aim} and {01918183-v walk leisurely — no gloss}. A single NomLex-PT link, then, allow us to find synsets that are good candidates to be merged, which seems work that some of the PWN team are interested in doing.

Another challenge is the fact that a single verb in Portuguese, like *esperar* can have two nominalizations *esperança* and *espera* that correspond to two different verbs in English, "hope" (with nominalization "hope") and "wait" (again with nominalization "wait").

Deciding which semantic type each pair in Portuguese should have is also a very difficult question, the differences between the types are far from obvious and many senses are not specific enough to be easily distinguished. Within our own group we have fierce discussions on the need or advantages of having all the many varieties of links of the Princeton WordNet as compared to our smaller set of underspecified labels from our previous work.

5. Implementing morphosemantic links

To construct the Portuguese morphosemantic links we combined the information from the NomLex-PT database with the Princeton morphosemantic links as follows (see the intuitive rule 1 below).

We look for a NomLex-PT relation between a verb in Portuguese $verb_{pt}$ and a noun in Portuguese $noun_{pt}$. Then we look for a morphosemantic link between a word sense of a verb in English $sense^{v}_{en}$ and a noun in English $sense^{n}_{en}$ (belonging to synsets ss^{1}_{en} and ss^{2}_{en} respectively) where $verb_{pt}$ appears in the list of words in ss^{1}_{pt} and $noun_{pt}$ appears in the lists of words in ss^{2}_{pt}. The synsets ss^{1}_{pt} and ss^{2}_{pt} are the Portuguese synsets related to the PWN synsets ss^{1}_{en} and ss^{2}_{en}, respectively.

We have a one-to-one relation between English synsets and Portuguese synsets, by construction. Because the relation between nouns and verbs in $nomlex(verb_{pt}, noun_{pt})$ was manually created in NomLex-PT and the semantic relation in $morpholink(sense^{v}_{en}, sense^{n}_{en}, type)$ was manually checked by the PWN team, the first two conjuncts are, in principle, true. The conjuncts relating synsets in PWN and OWN-PT are true by the alignment of the wordnets. So we have to make sure that the senses in both PWN and the OWN-PT are correct, to infer a correct morphosemantic link in Portuguese.

$$\begin{aligned} morpholink(sense^{v}_{en}, sense^{n}_{en}, type) \\ \wedge\ nomlex(verb_{pt}, noun_{pt}) \\ \wedge\ sense(sense^{v}_{en}, verb_{en}, ss^{1}_{en}) \\ \wedge\ sense(sense^{n}_{en}, noun_{en}, ss^{2}_{en}) \\ \wedge\ same(ss^{1}_{en}, ss^{1}_{pt}) \wedge\ same(ss^{2}_{en}, ss^{2}_{pt}) \\ \wedge\ sense(sense^{v}_{pt}, verb_{pt}, ss^{1}_{pt}) \\ \wedge\ sense(sense^{n}_{pt}, noun_{pt}, ss^{2}_{pt}) \\ \rightarrow morpholink(sense^{v}_{pt}, sense^{n}_{pt}, type) \end{aligned} \quad (1)$$

Figure 1 shows an example of a Portuguese morphosemantic link between to senses of the words *cantor* and *cantar* produced by rule 1 from the English morphosemantic link between the senses of the words "singer" and "sing". In Figure 1 note also the bidirectional link between senses called "derivationallyRelated" from PWN. Note that all morphosemantic links are among noun and verb synsets containing words that share an underlying meaning and are derivationally related.

From the heuristics above, we got 2,735 possible links to add to OpenWordNet-PT. Some of them are mistakes that came from the original wordnet directly, as the pair "*confess-confessor*" linked by an *agent* link. The nominal "confessor", the priest that hears the confession, is not the agent of the confession, but its *patient/undergoer*, usually. This is the same kind of issue as the "*suspect-suspect*" above. Those cases will be reported and hopefully discussed with the PWN team.

Some more mistakes were found when looking at the Portuguese senses and we are in the process of removing these. However, this is a laborious project, because, as we mentioned before, we want to use the incorrect links as opportunities to complete the correct ones.

When we cannot find all the elements in rule 1 above, say we miss a Portuguese verb sense, but the other elements are present, we can use SPARQL queries to produce large numbers of possible candidate links. Clever heuristics, perhaps using the number of senses of a verb or using words in glosses and definitions might be used to narrow down the number of candidate links reasonable to present to the evaluation team. For example the algorithm produced the pair *bray-bray* (*zurrar-zurro*) for which we did not have the verb *zurrar* in OWN-PT. (Actually we had the word only in the "wrong" or metaphoric meaning of "bray", that is to laugh coarsely.) The OpenWordNet-PT missed the literal meaning of *making the noise characteristic of donkeys*. But both NomLex-PT and the morphosemantic links had the connection, so it was suggested and we could add it. In

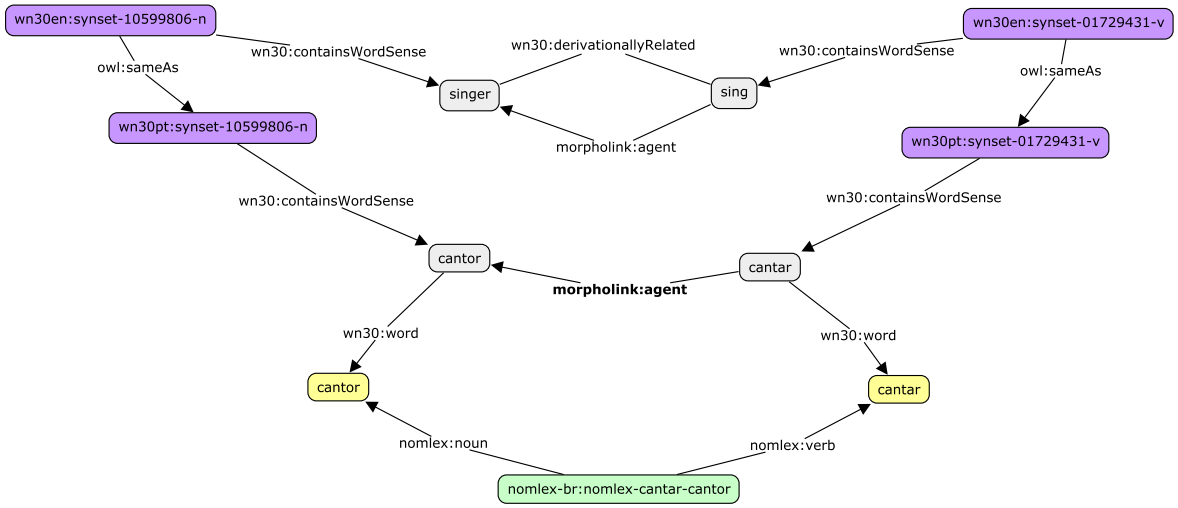

Figure 1: Subgraph showing one projection of an English morphosemantic link to Portuguese (bold). The gray nodes are Word instances, the yellow nodes are the WordSense instances and synsets are the purple ones. Remember the a word may have more than one sense, being each sense connected to at most one synset.

our GitHub repository we have listed those candidate lists and while we can see that many will work as morphosemantic links in Portuguese we are still formulating heuristics to extract the most probable pairs to manually verify.

It is also true that our processing of parallel English and Portuguese links has allowed us to produce other collections of candidate links[3], that would get us closer to the almost 17k links that Princeton WordNet has, however one should remember that our Portuguese wordnet is roughly half the size of PWN.

6. Future Work

Improving an automatically created resource, to make sure that meanings are not mangled and that bad translations are not solidified, is a hard task. Even if our OpenWordNet-PT were to be fully manually verified, mistakes and omissions have a tendency to creep into big lexica and we all know that even the gold standard resources like PWN have failings. Some of them are failings of sparsity of linking between synsets, which the project on morphosemantic links was supposed to alleviate. Other failings are the too fine-grained character of some synsets that the Global WordNet Association seems to have decided to improve on, using a new collection of interlingual indices. In any case adding morphosemantic links to the exposed face of PWN and to OpenWordNet-PT seemed to us useful in many ways.

Despite the relatively lower numbers of links already added in comparison to the previous PWN numbers, the exercise of adding the morphosemantic links, verifying the ones automatically created and brainstorming on the ways of extending those to other links that we believe are missing in the Portuguese wordnet, but present in the English one, all seem valuable ways of improving the quality of our resource. Concretely, this exercise helped us to address many issues in OpenWordnet-PT, removing and adding words from synsets to make the new links work. While trying to complete this process, we need to consider ways of evaluating the improvement that we are achieving. This is another hard problem that we expect to address soon.

As further future work we realize that the work relating nouns and verbs described in the creation of NomLex-PT is just a beginning. Similar work needs to be done to relate semantically verbs and adjectives *redden-red* (*avermelhar-vermelho*) and pairs of adjectives and adverbs *fast-fast* (*rápido-rapidamente*). Also a serious discussion on what are exactly the morphosemantic links we need for specialized tasks is required. The hope is to break down this work in small chunks and to be able to build on other researchers' work, if their research is open source and open use.

[3]The reports of missing nouns and verbs are available at `http://wnpt.brlcloud.com/wn/prototypes`.

7. Bibliographical References

Alexiadou, A. (2001). *Functional Structure in Nominals: Nominalization and Ergativity*. John Benjamins Publishing.

Asher, N. (2011). *Lexical Meaning in Context. A Web of Words*. Cambridge University Press, Cambridge.

Balvet, A., Barque, L., Condette, M., Haas, P., Huyghe, R., Marín, R., and Merlo, A. (2011). La ressource Nomage. *TAL*, 52(3):1–24.

Bilgin, O., Cetinoglu, O., and Oflazer, K. (2004). Morphosemantic relations in and across wordnets. In *Global WordNet Conference*. Global WordNet Association.

Brandtner, R. (2011). Deverbal nominals in context: Meaning variation and copredication. *SinSpeC*, 8.

Chomsky, N. (1970). Remarks on Nominalization. In R. A. Jacobs et al., editors, *Readings in English Transformational Grammar*. Waltham, Massachussets.

Clark, Eve.; Clark, H. (1979). When nouns surface as verbs. *Language*, 55(4):767–811, dec.

Costa, L., Santos, D., and Rocha, P. A. (2009). Estudando o português tal como é usado: o serviço AC/DC. In *The 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009)*, 8-11 de Setembro.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *Proc. of 24th International Conference on Computational Linguistics*, COLING (Demo Paper).

de Paiva, V., Real, L., Rademaker, A., and de Melo, G. (2014). NomLex-PT: A Lexicon of Portuguese Nominalizations. In N. Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Fellbaum, C.; Osherson, A. and Clark, P. (2007). Putting semantics into WordNet's "Morphosemantic" links. *Proc. of the Third Language and Technology Conference, Poznan, Poland. Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies, eds. Z. Vetulani and H. Uszkoreit. Springer Lecture Notes in Informatics vol. 5603:350-358 (2009)*.

Gurevich, O. and Waterman, S. A. (2009). Mining of parsed data to derive deverbal argument structure. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks*, GEAF '09, pages 19–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gurevich, O., Crouch, R. S., King, T. H., and Paiva, V. D. (2008). Deverbal Nouns in Knowledge Representation. *Journal of Logic and Computation*, 18:385–404.

Habash, N. and Dorr, B. (2003). A categorial variation database for english. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 17–23. Association for Computational Linguistics.

Jezek, E. and Melloni, C. (2009). Complex types in the (morphologically) complex lexicon. *Proceedings of the GL2009*.

Jezek, E. and Melloni, C. (2011). Nominals, Polysemy, and Co-predication. *Journal of Cognitive Science*, 22:1–31.

Macleod, C., Grishman, R., Meyers, A., Barret, L., and Reeves, R. (1998). Nomlex: a lexicon of nominalizations. *Proceedings of Euralex*.

Peris, A. and Taulé, M. (2011). AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural*, 46:11–18.

Pustejovsky, J. (1995). *The generative Lexicon*. MIT Press, Crambridge.

Real, L. and Retoré, C. (2014). On the semantics of deverbals in a richly typed system. *Journal of Logic, Language and Information*.

Stoyanova, I., Koeva, S., and Leseva, S. (2013). Wordnet-based cross-language identification of semantic relations. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 119–128, Sofia, Bulgaria, August. Association for Computational Linguistics.