Proceedings

Workshop on Corpora and Tools for Processing Corpora

Collocated with PROPOR 2016 – The 12th International Conference on the Computational Processing of Portuguese

July 13, 2016
Tomar, Portugal

Hilário Leal Fontes and António Branco (eds.)

CM2News: Towards a Corpus for Multilingual Multi-Document Summarization

Ariani Di-Felippo[1,2]

[1] Interinstitutional Center for Computational Linguistics (NILC), São Carlos/SP, Brazil
Av. Trabalhador São-carlense, 400, São Carlos, 13566-590, Brazil
[2] Language and Literature Department, Federal University of São Carlos (UFSCar)
Rodovia Washington Luís, km 235 - SP 310, São Carlos, 13565-905, Brazil
arianidf@gmail.com

Abstract. This paper describes the ongoing construction of CM2News, a semantic-annotated corpus for fostering research on multilingual multi-document summarization. The corpus comprises 20 clusters of news texts in English and Brazilian Portuguese languages and a set of multi-document manual and automatic summaries. All the source texts have a layer of semantic annotation at lexical level. Some clusters also have annotation at sentence level, as well as alignment of texts and human summaries. The corpus is a result delivered within the context of the *Sustento* Project, which aims at generating linguistic knowledge for multi-document summarization. The corpus design and the manual annotation tasks are detailed in this paper.

Keywords: corpus, multilingual resource, multi-document summarization.

1 Introduction

As the amount of on-line news texts in different languages is growing at an exponential pace, Multilingual Multi-Document Summarization (MMDS) is a quite desirable task. It aims at identifying the main information in a cluster of (at least) two news texts, one in the user's language and one in a foreign language, and presenting it as a coherent/cohesive summary in the user's languages.

The ongoing Sustento project[1] tackles this and also other multi-document summarization tasks. Specifically, it has been focusing on 3 correlated tasks: (i) characterization of the human multi-document summarization (HMS) and development of automatic methods based on HMS strategies, (ii) study of the multi-document phenomena (e.g., redundancy) and proposition of methods for their automatic detection, and (iii) development of deep methods based on semantic-conceptual representation of the source texts. The project is mainly corpus-driven, i.e., linguistic descriptions, tools and applications are drawn upon corpora. This motivates our interest in constructing CM2News, a *Multi-document Bilingual Corpus of News Texts* for MMDS, which was first described in [1].

[1] http://www.nilc.icmc.usp.br/nilc/index.php/team?id=23

The CM2News comprises 20 clusters of news texts. Each cluster is composed of 2 source-texts, 1 in English (En) and 1 Brazilian Portuguese (BP), and a set of multi-document manual and automatic summaries. Given our interest in exploring deep summarization based on semantic-conceptual knowledge, all the source texts were manually annotated using Princeton WordNet [2], and some clusters were also annotated following UNL (*Universal Network Language*) formalism [3]. We also carried out the sentential alignment of texts and human summaries of some clusters based on overlapping content between the sentences.

To the best of our knowledge, CM2News is the first multi-document corpus with multilingual clusters that include Portuguese. This paper focuses on its manual annotation in order to produce a resource for MMDS. Section 2 first reports the corpus design. Section 3 focuses on the annotation tasks, which include the meaning representation following two different conceptual models, and the alignment of texts and human summaries. In Section 4, we briefly highlight the projects that already made use of CM2News. Section 5 provides some final remarks and future works.

2 Building Principles

According to [4], a well-designed corpus should reflect its purpose. Since our corpus has been building for MMDS, it is a multi-document and multilingual resource. This means that its internal structure is based on *clusters*, and each cluster is composed of texts in different languages on the same topic. The CM2News corpus has 20 clusters, and each of them is composed of 2 news texts, 1 in En and 1 in BP. The corpus sums up 40 texts altogether, amounting to 19.984 words.

The texts in En and BP were manually collected from the *BBC*[2] and *Folha de São Paulo*[3] on-line news agencies, respectively. To collect them, we have followed the 3 criteria that were applied to build CSTNews [5], a reference corpus in BP for Multi-Document Summarization (MDS). One criterion was to collect texts with similar length (in terms of words). For example, the texts in En and BP of the cluster C19 have 446 and 452 words, respectively. Another criterion was selecting topics with high popularity on the web, which means that CM2News only cover trending topics at the time of the corpus construction (e.g., "Angelina Jolie's mastectomy" in 2013). Finally, according to the diversity guideline, the clusters cover a variety of domains, i.e., world (8 clusters), politics (3 clusters), health (4 clusters), science (3 clusters), entertainment (1 clusters), and environment (1 clusters). Moreover, each cluster of our corpus also has 1 human multi-document *abstract*[4], and automatic multi-document extracts generated by baseline and deep MMDS methods. Both human and automatic summaries are written in Portuguese, but they are ideally brief representations of the essential content of the two source-texts. All summaries were generated based on a compression rate of 70%, which means that they correspond to 30% of the size of the largest text of the cluster.

The next section describes the linguistic annotations of the CM2News corpus.

[2] http://www.bbc.co.uk/news/

[3] http://www.folha.uol.com.br/

[4] Abstracts are summaries that contain some degree of paraphrase of the input.

3 Corpus Annotation

3.1 Lexical Semantic Annotation

The 40 source texts of the CM2News corpus have a layer of semantic annotation at lexical level. Specifically, the common nouns, which cover part of the main content of a text, were semi-automatically tagged with their correspondent concept.

In order to identify the nominal concepts in the texts, we made use of WordNet[5] lexical database. Although WordNet's fine-grained senses may create difficulties for annotating nouns, we have chosen such database due to its widespread application in several NLP tasks and broad coverage, and the still partial development of similar resources for Portuguese.

The annotation was carried out by groups of 2 or 3 experts[6], in a total of 12 computational linguists, in daily meetings from 90 to 120 minutes. The annotation process, including 1 day for training, took the period of 15 days. For each cluster to annotate, the experts were organized in different groups, trying to avoid any annotation bias. To assist the experts, we built an easy-to-use editor called MulSen[7] (Multilingual Sense Estimator). Given a cluster, the editor first performs an automatic pre-processing step, which consists in annotating the source texts with part-of-speech (POS) tags. MulSen incorporates two taggers, one for each language, and the output of such tools can be manually revised if necessary. Once the texts are tagged, the annotation of a noun *n* in Portuguese, in particular, starts with the automatic translation of *n* to English, since WordNet codifies the concepts by sets of synonyms in English. The translation is performed using the online bilingual dictionary WordReference®[8], but the editor also allows the manual inclusion of a translation equivalence. Finally, the editor suggests the best synset that represents the underlying concept of *n,* which should be validated by the experts to complete the process. The suggestion results from the application of a word sense desambiguation algorithm [6]. If the suggested synset is not appropriate, the editor displays all the synsets containing the English translation of *n* and then the annotators are able to select a more suitable option among them. The annotation of a text in English basically follows the same procedure except the machine-translation stage.

The experts have followed 4 general rules in order to annotate the nouns: (i) firstly annotate the text in English of a cluster, since its vocabulary can provide appropriate translations for the annotation of the nouns in Portuguese, (ii) annotate the POS silence, i.e., nouns that were not automatically detected, (iii) ignore the POS noise, i.e., words that were wrongly annotated as nouns, and (iv) annotate every occurrence of a concept (i.e, synonyms and equivalences) in the cluster with the same (and more adequate) synset.

---

[5] A semantic network of English in which the meanings of words and expressions of noun, verb, adjective, and adverb classes are organized into "sets of synonyms" (*synsets*). Each *synset* expresses a distinct concept and they are interlinked through conceptual-semantic (hyponymy, meronymy, entailment, and cause) and lexical (antonymy) relations [2].

[6] The agreement rate has not been calculated yet.

[7] http://www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html

[8] http://www.wordreference.com/

The annotation was also performed according to 4 specific rules. Since the taggers only detect single word forms, the first rule establishes that every common noun that is a multiword expression head should be annotated with a synset that codifies the expression's sense. For instance, the head (shown in italics) of the multiword expression "*gás* de pimenta" ("pepper spray") was annotated with the *synset* {pepper spray} ("a nonlethal aerosol spray made with the pepper derivative oleoresin capiscum"). Following this rule, we were able to encode complex concepts by annotating single words only. The second rule determines that the annotators should analyze all the possible translations provided by WordReference before selecting one. This is particularly important because the adequate translation may not be the first option in the list of equivalences provided by the editor. The same procedure should be followed regarding the synset selection. When the editor suggests an inadequate synset, the annotators should carefully analyze the other options retrieved from the database. For cases where translations have to be manually inserted in the editor, the third rule establishes that the annotators should look for equivalences in external resources (e.g., *Google Translator*®[9], *Linguee*[10], and other dictionaries) and analyzes all synsets retrieved from WordNet by testing the equivalences. The forth rule determines that, if a specific concept is not covered by WordNet, it should be selected a more general one. This means that, if any of the synsets retrieved by the chosen translation is adequate, the annotators should look for a satisfactory hypernym synset.

The example (1) provides an illustration of an annotated sentence. The 4 nouns (shown in bold) that occur in the English sentence "Brazil's opening Confederations Cup match was affected by protests that left 39 people injured" (C17) were tagged with the correspondent synset, indicated between "{}". For a better comprehension, we provide the gloss (i.e., an information definition of the concept) of each synset.

(1) Brazil's **opening**<{opening} "a ceremony accompanying the start of some enterprise"> Confederations Cup **match**<{match} "a formal contest in which two or more persons or teams compete"> was affected by **protesters**<{dissenter, dissident, protester, objector, contestant} "a person who dissents from some established policy"> that left 39 **people**<{people} "any group of human beings"> injured.

3.2 Sentential Semantic Annotation

Besides the semantic annotation at lexical level, some clusters were also annotated at sentential level[11], a task first described by [7]. Both source texts and human summaries were annotated with the UNL [10] formalism, in a process called UNLization. UNL is aimed at expressing information conveyed by natural language (NL) sentences through binary relations between concepts [7]. Thus, UNL is not different from the other formal languages devised to represent NL sentence meaning [8]. The general syntax of the relations is RL(UW1,UW2), where RL stands for a Relation Label, which signals the semantic relation, and UWn, for Universal Words, which signal the related concepts. RLs are specified through mnemonics, for example, *agt* for *agent*, *mod* for *modifier*, or *obj* for *object*. UWs, in particular, constitute the

[9] https://translate.google.com/
[10] http://www.linguee.com/
[11] There is no connection between the lexical and sentential annotations so far.

UNL vocabulary, and can be annotated by attributes to provide further information on the circumstances under which they are used (e.g., tense and aspect). Those are signaled by Attribute Labels (ALs). According to [9], the advantages of UNL are: (i) flexibility and neutrality, since it is a language created to represent any content in any domain in any language, and (ii) generality, since the set of UWs[12] and *RLs* is sufficient to describe any kind of content expressed in NLs.

From the 20 clusters, three (C1, C2, and C9) were annotated with UNL, in a total of 158 sentences (3504 words). Each cluster was manually tagged with the support of the UNL Editor [10]. One computational linguist carried out the task in two-hours daily sessions, during 3 months. Given a text, the editor first split it into sentences and then the UNLization follows 3 stages: (i) identification of concepts (Stage 1); (ii) assigning attributes (Stage 2), and (iii) identification of relations between concepts (Stage 3). The UNLization of the English sentence "*Seven people have been rescued from the rubble*" is shown in Figure 1. In Stage 1, we identified 4 UWs making use of the dictionaries available in the editor: "7", "person", "rescue", and "rubble". In Stage 2, the UW "person" received the attribute label "@pl", which means that there is more than one person (plural) involved in the event. The UW "rescue" has two ALs: "@past", which indicates that the event took place in the past, and "@entry", which means that this is the main UW of the sentence. The UW "rubble" received the attribute "@def", which expresses definiteness. In Stage 3, three RLs were identified: "qua" (quantity), "obj" (affected thing), and "src" (source). The binary RL "qua", for example, interconnects the UWs "7" and "person". Next, we describe the manual alignment of source texts and human summaries.

Stage 1	Stage 2	Stage 3
7 person rescue rubble	7 person.@pl rescue.@past.@entry rubble.@def	*qua*(person.@pl,7) *obj*(rescue.@past.@entry,person.@pl) *src*(rescue.@past.@entry,rubble.@def)

Fig. 1. Sentence UNL encoding.

3.3 Alignment of Abstracts and News Texts

Many authors have been using manual alignment of texts and reference summaries in Automatic Summarization, since it may reveal some of the human strategies used to produce the summary [11], [12]. Thus, one computational linguist has performed the alignment in one-hour daily sessions, during 1 month. The expert has followed the methodology described in [13] to align 3 clusters (C1, C2 and C9). The manual alignment was performed in the summary-to-documents direction and at sentence level, and the links were established based on total or partial content overlap. In this multi-document setting, a summary sentence may be aligned to more than one document sentences. Once the raw sentences were linked, their correspondent UNL codifications were also connected. Figure 2 illustrates the alignment.

[12] Although UWs take their meanings from English word senses, each universal word expresses a very definite meaning so lexical ambiguity is kept to a minimum.

Summary sentence / UNL codification	Source sentence / UNL codification
Cerca de 100 pacientes tiveram que ser retirados do centro médico. (*About 100 patients had to be removed from the medical center*) [C9_S2]	Nearly 100 patients at the St John Regional Medical Center in Joplin were evacuated after the hospital took a direct hit. [C9_En_S30]
	Pacientes tiveram que ser retirados do centro médico. (*Patients had to be withdrawn from the medical center*) [C9_BP_S9]
obj(remove.@past.@obligation.@entry,patient.@pl) mod(center.@def,medical) src(remove.@past.@obligation.@entry,center.@def) qua(patient.@pl, nearly) bas(nearly,100)	bas(nearly,100) qua(patient.@pl,nearly) plc(patient.@pl,St John Regional Medical Center.@def) plc(St John Regional Medical Center.@def,Joplin) obj(evacuate.@past.@entry,patient.@pl) tim(evacuate.@past.@entry,after) obj(after,01) aoj:01(direct,hit.@indef) obj:01(take.@past.@entry,hospital.@def) agt:01(take.@past.@entry,hit.@indef)
	obj(remove.@past.@obligation.@entry,patient.@pl) mod(center.@def,medical) src(remove.@past.@obligation.@entry,center.@def)

Fig. 2. Alignment of sentences and their correspondent UNL encodings.

In Figure 2, for example, the summary sentence S2 is aligned to the following two source sentences because they share the main information: S30 from the English text and S9 from the Portuguese text. Thus, their correspondent UNL representations were linked as well. Table 1 shows the distribution of the different alignment types (1-*n*). Table 2 describes the number of alignments where a summary sentence was aligned to source sentences(s) in just one language (Portuguese or English) or in both languages.

Table 1. Distribution of the alignment types in the corpus.

Alignment	1:1	1:2	1:3	1:4	1:5	1:6	1:7	1:8	1:9	1:10
Quantity	8	7	4	0	3	0	0	0	0	1

Table 2. Distribution of the alignments per language.

Alignment	Summary:Portuguese	Summary:English	Summary:Both
Quantity	6	6	11

According to the results, we may see that 8 summary sentences were aligned to only one sentence of the source texts (1-1), 7 summary sentences were aligned to 2 sentences of the source texts (1-2), and so on. The alignment illustrated in Figure 2, for example, is 1-2. From the 23 summary sentences, 15 were aligned (65,3%) to some source sentence, with the distribution per language as described in Table 2. This result was expected, since a multi-document summary could be potentially connected to 2 related source texts of its cluster. From the 144 sentences in the source texts, 50 (37,4%) were aligned to some summary sentence, but it does not mean that the sentences were aligned only once. A sentence of a summary may be aligned to more than one sentence of the source text, and the sentences of the source texts may be redundant or even identical. Next, we give an idea on how the CM2News corpus has been used in MMDS.

4 CM2News in MMDS Projects

Using CM2News, [1] has developed two deep MMDS methods for generating extracts in Portuguese. The methods select sentences to compose extracts based on the frequency of occurrence of their nominal concepts in the cluster. To score and rank the sentences, they make use of the synset annotation. The CF (concept frequency) method selects the top-ranked sentences, independently of their source language. If a selected sentence is in English, it is automatically translated to Portuguese. The CFUL (concept frequency + user language) method is driven by the user's language. It exclusively selects the top-ranked sentences from the text written in Portuguese to compose the summary, also avoiding redundancy. In an intrinsic evaluation, the methods have outperformed a sentence position *baseline* (which applies a MT strategy over the source texts) in terms of informativiness and linguistic quality.

Using the UWs from the UNL annotation, [7] has explored 3 conceptual measures to capture relevant content in MMDS: (i) CF (concept frequency), (ii) CF*IDF (concept frequency corrected by the inverted document frequency), and (iii) CF/No. of Cs (concept frequency normalized by the number de concepts in the sentence). The author has compared the measures to a superficial *sentence position* method. To evaluate the potential of the measures in capturing human preferences, the author ranked the source sentences according to each strategy, and calculated how many aligned source-sentences were covered by the top sentences of each rank. The concept-based method with the best performance is (iii), but it does not outperform the sentence position method.

4 Future Works and Final Remarks

This paper described the linguistic annotation of the CM2News corpus, which aims at supporting the investigation of deep strategies on MMDS involving Portuguese. The corpus and tools are all available on the *Sustento* Project website. We hope CM2News may foster research not only on summarization and semantic analysis, but also in other Natural Language Processing areas. Future work includes increasing the quantity of clusters, extending the UNL annotation to the entire corpus, and annotating other kinds of lexical concepts, as those expressed by verbs, for example.

Acknowledgments. We thank CNPq (#483231/2012-6), and FAPESP (#2012/13246-5) for the financial support.

References

1. Tosta, F.E.S.: Aplicação de conhecimento léxico-conceitual na Sumarização Multidocumento Multilíngue. 2013. Dissertação (Mestrado em Linguística) - Departamento de Letras, Universidade Federal de São Carlos (2014)
2. Fellbaum, C. (Ed.): Wordnet: an electronic lexical database (Language, speech and communication). Massachusetts: MIT Press (1998)
3. Uchida, H., Zhu, M.; Della Senta, T.: The UNL, a Gift for a Millennium. The United Nations University - Institute of Advanced Studies, Tokyo, Japan (1999)
4. Sinclair, J.: Corpus and Text - Basic Principles. In: Wynne, M. (Ed.). Developing Linguistic Corpora: a Guide to Good Practice, Oxbow Books, pp. 1-16 (2005)
5. Cardoso, P.C.F., Maziero, E.G., Jorge, M.L.C., Seno, E.M.R., Di-Felippo, A., Rino, L.H.M., Nunes, M.G.V., Pardo, T.A.S.: CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: 3rd RST Brazilian Meeting, pp. 88-105. Cuiabá, MT, Brazil (2011)
6. Nóbrega, F.A.A.: Desambiguação lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - ICMC, USP, São Carlos (2013)
7. Chaud, Matheus. R. Investigação de estratégias de seleção de conteúdo baseadas na UNL (*Universal Networking Language*). 2015. 157 f. Dissertação (Mestrado em Linguística) - Universidade Federal de São Carlos, São Carlos, SP, 2015.
8. Martins, R. T. et al.: The UNL distinctive features: evidences through a NL-UNL encoding task. In: 1rst International Workshop on UNL, other Interlinguas and their Applications, LREC, 2002. Las Palmas, pp. 08-13. (2002)
9. Cardeñosa, J. et al.: A new knowledge representation model to support multilingual ontologies. A case study. In: International Conference on Semantic Web and Web Services (SWWS), pp. 313-319. Springer Berlin Heidelberg, Berlin (2008)
10. Alansary, S., Nagi, M., Adly, N.: UNL Editor: An annotation tool for semantic analysis. In: 11th International Conference on Language Engineering. Cairo, Egypt (2011)
11. Marcu, D.: The automatic construction of large-scale corpora for summarization research. In: 22th Conference on Research and Development in IR, pp.137-44 (1999)
12. Hirao, T., Suzuki, J., Isozaki, H., Maeda, E.: Dependency-based Sentence Alignment for Multiple Document Summarization. In: International Conference on Computational Linguistics (COLING). Switzerland, pp. 446-452 (2004)
13. Camargo, R.T., Di Felippo, A., Pardo, T.A.S.: On Strategies of Human Multi-Document Summarization. In: 10th Brazilian Symposium in Information and Human Language Technology - STIL, pp. 141-150. Natal, Brazil (2015)

Language Resources and Processing Tools at the University of Lisbon in the NLX Group Collection

António Branco, João Silva, Francisco Costa, João Rodrigues, Pedro Martins, Eduardo Ferreira, Filipe Nunes, Sérgio Castro, Steven Neale, Catarina Carvalheiro, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Andreia Querido, Rita de Carvalho, Marisa Campos, Nuno Rendeiro, Catarina Correia, Patrícia Gomes, Diana Amaral, and Rita Valadas Pereira

University of Lisbon, Faculty of Sciences, Department of Informatics

Abstract. In this paper we present many of the language resources and processing tools developed and made available at the University of Lisbon by the NLX - Natural Language and Speech Group. These were developed over the years to support the development of a wide array of natural language applications, including machine translation.

Keywords: Portuguese, language technology, natural language processing, language resources, language processing tools.

1 Introduction

The development of machine translation solutions requires a number of instrumental and auxiliary language processing tools as well as appropriate companion data sets for the training and evaluation of these tools and applications. This paper aims at providing a brief introduction to the collection of processing tools and language resources for the Portuguese language developed and made available at the University of Lisbon by the NLX Group, the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon.

These resources and processing tools are made available from the NLX-Group website.[1] Most of them support also free online linguistic processing services and demos that are available at the LX-Center.[2]

This paper is organized as follows: Section 2 presents the collection of tools that are instrumental for natural language processing and machine translation, and Section 3 covers the language resources. The paper closes with final remarks in Section 4.

[1] http://nlx.di.fc.ul.pt/
[2] http://lxcenter.di.fc.ul.pt/

2 Language processing tools in the NLX Collection

In the NLX-Group we have developed language processing tools that virtually cover the full range of tasks from shallow to deep processing. Some of tools address simple procedures, such as the LX-Tokenizer, and others tackle more sophisticated functionalities, such as the Lx-DepParser, and others cover named entity recognition, verbal and nominal inflection or word-sense disambiguation, etc. These tools are useful at different levels to support machine translation and are presented below.

- **LX-Lemmatizer** is a verbal lemmatizer that takes a Portuguese verb form as input and delivers a ranked list of the corresponding lemmata (infinitive forms) together with inflectional feature values[18]. Its performance was evaluated as delivering 96.5% accuracy.
- **LX-Inflector** is a language processing tool for nominal lemmatization and inflection [4], taking a Portuguese word form that follows the nominal inflection paradigm and an inflection feature bundle, and delivers both the corresponding lemma and the indication of its feature bundle, and the resulting form that conveys the feature bundle entered. It is based on principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, thus handling neologisms. The lemmatization function has 97.7% f-score. Figure 1 shows LX-Inflector online service.

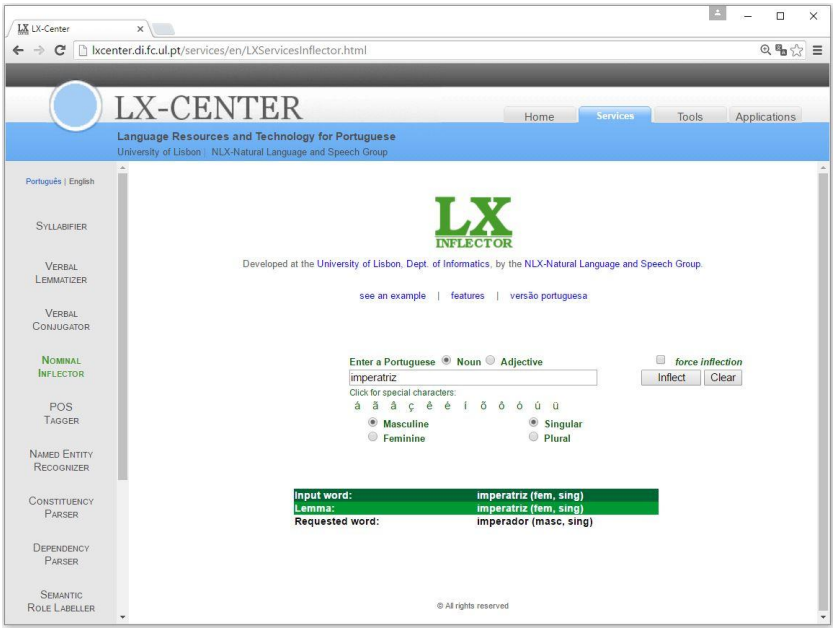


Fig. 1. LX-Inflector online service for the nominal lemmatization and inflection of Portuguese

- **LX-Chunker** is an identifier of paragraphs and sentences for Portuguese. It seeks to cope with the ambiguity and ambivalence of symbols that in some occurrences are indicators of separations among sentences and in other contexts are not. It is a hybrid tool, based on regular expressions and hidden Markov models, with an f-score of 99.9%.
- **LX-Conjugator** is a verbal conjugator for the Portuguese language [18]. It takes a Portuguese infinitive verb form as input and delivers the corresponding conjugated forms. It is the only available tool for fully-fledged Portuguese verb conjugation, including the full range of pronominal conjugation forms. Its capacity includes the handling of pronominal conjugation, compound tenses, double forms of past participles, past participle forms inflected for number and gender, negative imperative forms, and courtesy forms for second person. Given that it is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, it is the only available conjugator to handle neologisms. Their occasional faults have been correct along the time as it has been put to use, and at present no defect is known.
- **LX-Tokenizer** is an identifier of the boundaries of relevant word-level tokens in Portuguese text. It seeks to cope with the ambiguity of strings that in some contexts are single-word tokens and in some other contexts are contractions, i.e. double-word tokens. It achieves an f-score of 99.7%. It is incorporated in Lx-Suite[7], available at the Lx-Center.[3]
- **LX-Tagger** is a part-of-speech tagger with disambiguation and full coverage for the Portuguese language. For each word occurring in a text and from the possible different morpho-syntactic categories that word may have in the lexicon, it assigns a single tag to it that indicates the morpho-syntactic category that it bears in that occurrence in the text. It scores 96.8% accuracy.
- **LX-NER** is an identifier and classifier of named entity expressions for the Portuguese language [9]. Its number-based part evaluates to an f-score of 85.6%, and the name-based to 85.7%.
- **LX-NED** is a named entity disambiguator that annotates the occurrence of an input expression with the Wikipedia entry it refers to in its context, with an f-score of 67.0%.
- **LX-WSD** is a word sense disambiguator that annotates the occurrence of an input word with the MWN.PT wordnet concept it expresses in its context, with an fscore of 65.0%.
- **LX-Parser** is a stochastic parser that performs the syntactic analysis of Portuguese sentences in terms of their constituency structure[17][16]. It achieves an f-score of 88% under the Parseval metric.
- **LX-DepParser** is a parser of grammatical dependency relations for sentences of Portuguese that for each input sentence delivers a graph connecting its words and whose directed arcs represent grammatical dependencies and the labels at the said arcs represent the grammatical function of those dependencies. The evaluation of its performance obtained 91.2% in terms

[3] http://lxcenter.di.fc.ul.pt/

of labelled attachment score (LAS). Figure 2 shows LX-DepParser online service.

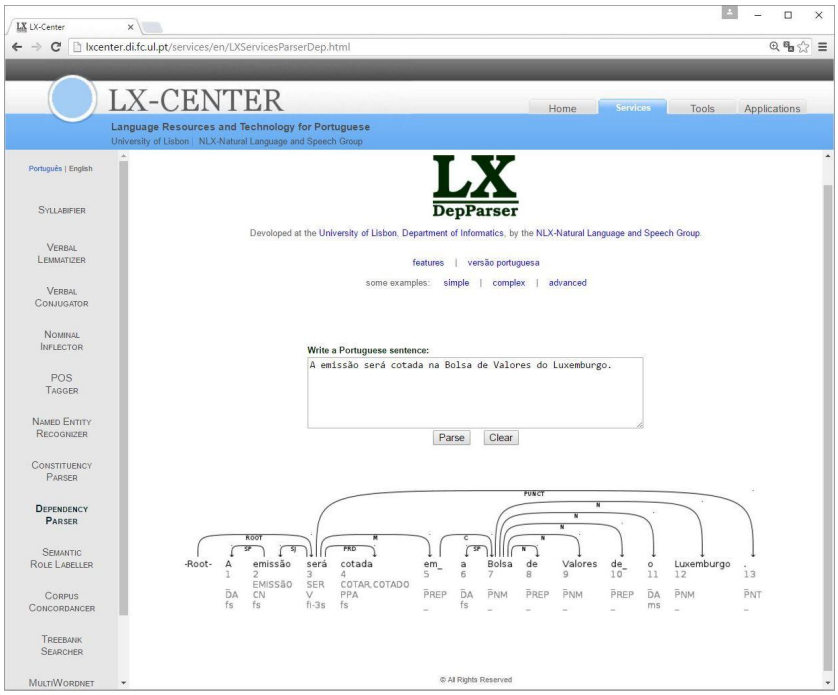


Fig. 2. LX-DepParser online service for the syntactic analysis of Portuguese

3 Language Resources in the NLX collection

In this section we briefly introduce language resources in the NLX-Group collection that are relevant for the theme of the present workshop.

- **CINTIL-International Corpus of Portuguese**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including morphological analyzers, POS taggers and named entity recognizers. This corpus contains 1 million words manually annotated by experts in natural language science and technology. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It was developed in cooperation with CLUL-Center of Linguistics of the University of Lisbon[1].
- **CINTIL-DeepBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars [15]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts

in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms.

- **CINTIL-Treebank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers[15]. This treebank contains around 10 000 sentences (100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree tagged with phrase categories. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-DependencyBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers[15]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions holding between its words[15]. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-PropBank**: Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers[3]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. The syntactic constituents of sentences are associated to linguistic information about its semantic role. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL-LogicalFormBank**: Set of text materials to support the evaluation and training of tools for the semantic processing of Portuguese[15][2]. This corpus contains around 10 000 sentences (approximately 100000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the logical form that represents its meaning in a logical language for semantic description[15].
- **CINTIL-WordSenses**: Set of text materials to support the evaluation and training of word sense disambiguators[13]. This corpus contains around 24 000 sentences with 45 000 words that are manually annotated by experts in natural language science and technology with the identifiers of concepts (synsets) that they convey in terms of the lexical semantic network MWN.PT [12]. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **CINTIL DependencyBank PREMIUM**: Set of text materials similar in design to the previous one and differing from it in the sentences that were treebanked and in the circumstance that the support tool to draw the grammatical dependency graphs is not the LXGram but the full text coverage

LXDependencyParser [8][6]. It contains 3 000 sentences (approximately 79 000 words).

- **CINTIL-NamedEntities**: Set of text materials to support the evaluation and training of named entity disambiguators. This corpus contains around 30 000 sentences with 26 000 named entities that are manually annotated by experts in natural language science and technology with identifiers of the corresponding entities in the DBpedia ontology[13]. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **QTLeap Multilingual Parallel Corpora**: Set of 4 000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software[11]. This textual material was collected using a commercial support service via chat, in Portuguese, and the corpus is thus composed by naturally occurring utterances produced by users while interacting with that service. Each question answer pair is translated into seven languages, other than Portuguese, namely Czech, Basque, Bulgarian, Dutch, English, German and Spanish.
- **QTLeap WSD/NED Multilingual Corpora**: Set of text materials comprising the QTLeap Multilingual Parallel Corpora and the Europarl multilingual corpora for the Czech (9.2 Million tokens), Basque (5.2 Million), Bulgarian (4.9 M), English (53 M), Portuguese (5.7 M) and Spanish (57.1M) languages, automatically annotated at multiple semantic levels by processing tools for tokenization, lemmatization, part-of-speech tagging, named-entity recognition and classification, named-entity disambiguation, word sense disambiguation and coreference resolution[14].
- **DeepBankPT**:Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars [10]. This corpus contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms.
- **TreebankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers. This treebank contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **PropBankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers. This corpus contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associ-

ated to its syntactic constituency tree decorated with semantic roles. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.

- **DependencyBankPT**: Set of text materials translated into Portuguese from the Penn Treebank to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers. This treebank contains around 3 500 sentences (approximately 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions holding between its words. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions.
- **LogicalFormBankPT**: Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the semantic processing of Portuguese. This corpus contains around 3 500 sentences (ca. 45000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the logical form that represents its meaning in a logical language for semantic description[5].

4 Final Remarks

The NLX Group has developed the language processing tools and resources briefly introduced above. These datasets and tools are distributed from the NLX-Group website or at the META-SHARE [4] distribution platform. They are made available with the goal of being of help for further research and progress of the computational processing of the Portuguese language.

References

1. Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M.F., Nunes, F., Silva, J.: Open resources and tools for the shallow processing of portuguese: the tagshare project. In: Proceedings of LREC 2006. Citeseer (2006)
2. Branco, A.: Logicalformbanks, the next generation of semantically annotated corpora: key issues in construction methodology. Recent Advances in Intelligent Information Systems, Exit, Warsaw pp. 3–11 (2009)
3. Branco, A., Carvalheiro, C., Pereira, S., Silveira, S., Silva, J., Castro, S., Graça, J.: A propbank for portuguese: the cintil-propbank. In: LREC. pp. 1516–1521 (2012)
4. Branco, A., Nunes, F.: Verb analysis in a highly inflective language with an mff algorithm. In: Computational Processing of the Portuguese Language, pp. 1–11. Springer (2012)
5. Branco, A., Silva, J., Gonçalves, P., Costa, F., Silveira, S., Del Gaudio, R., Rodrigues, J., Castro, S., Rodrigues, L., Martins, P., et al.: The cintil and lx companion collections of language resources and tools for portuguese

[4] http://metashare.metanet4u.eu

6. Branco, A., Silva, J., Querido, A., Carvalho, R.: Cintil dependencybank premium handbook: Design options for the representation of grammatical dependencies (2015)
7. Branco, A., Silva, J.R.: A suite of shallow processing tools for portuguese: Lx-suite. In: Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations. pp. 179–182. Association for Computational Linguistics (2006)
8. de Carvalho, R., Querido, A., Campos, M., Valadas Pereira, R., Silva, J., Branco, A., in Press: Cintil dependencybank premium: A corpus of grammatical dependencies for portuguese. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
9. Ferreira, E., Balsa, J., Branco, A.: Combining rule-based and statistical methods for named entity recognition in portuguese. In: Actas da 5a Workshop em Tecnologias da Informaçao e da Linguagem Humana (2007)
10. Flickinger, D., Kordoni, V., Zhang, Y., Branco, A., Simov, K., Osenova, P., Carvalheiro, C., Costa, F., Castro, S.: Pardeepbank: Multiple parallel deep treebanking. Proceedings of TLT-2012, Lisbon, Portugal pp. 97–108 (2012)
11. Gaudio, R.D., Burchardt, A., Branco, A., in Press: Evaluating machine translation in a usage scenario. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
12. MultiWordNet: The MultiWordNet project. http://multiwordnet.fbk.eu/english/home.php (nd), accessed: 2015-01-13
13. Neale, S., Valadas Pereira, R., Silva, J., Branco, A., in Press: Lexical semantics annotation for enriched portuguese corpora. In: Springer (ed.) Lecture Notes in Artificial Intelligence
14. Otegi, A., Aranberri, N., Branco, A., Hajič, J., Popel, M., Simov, K., Agirre, E., in Press: Qtleap wsd/ned corpora: Semantic annotation of parallel corpora in six languages. In: Proceedings, LREC2016 - 10th Language Resources and Evaluation Conference (May 23-28 2016)
15. Silva, J., Branco, A., Castro, S., Costa, F.: Deep, consistent and also useful: Extracting vistas from deep corpora for shallower tasks. In: Proceedings of the Workshop on Advanced Treebanking at the 8th Language Resources and Evaluation Conference (LREC). pp. 45–52 (2012)
16. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-box robust parsing of portuguese. In: Computational Processing of the Portuguese Language, pp. 75–85. Springer (2010)
17. Silva, J., Branco, A., Gonçalves, P.N.: Top-performing robust constituency parsing of portuguese: Freely available in as many ways as you can get it. In: LREC (2010)
18. da Silva, J.R.M.F.: Shallow processing of portuguese: From sentence chunking to nominal lemmatization (2007)

Language resources for information extraction and semantic computing - NLP at PUCRS

Renata Vieira[1], Daniela do Amaral[1], Sandra Collovini[1], Evandro Fonseca[1], Artur Freitas[1], Larissa Freitas[1], Roger Granada[1], Lucas Hilgert[1], Lucelene Lopes[1], Daniela Schmidt[1], Bernardo Severo[1], Marlo Souza[1], Cassia Trojahn[2]

[1] PUCRS - Porto Alegre, Brazil
{renata.vieira,lucelene.lopes}@pucrs.br
{daniela.amaral,sandra.abreu,evandro.fonseca,artur.freitas,
larissa.freitas,roger.granada,lucas.hilgert,daniela.schmidt,
bernardo.severo,marlo.souza}@acad.pucrs.br
[2] UT2J & IRIT - Toulouse, France
cassia.trojahn@irit.fr

Abstract. In this paper we present an overview of the language resources developed at the Natural Language Processing Lab at PUCRS, making them available to the research community.

Keywords: Information extraction, semantic computing, language resources

1 Introduction

At PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul) NLP Lab we investigate several semantic information extraction related problems, for which we have used and developed a series of corpora, annotations and tools. Our main themes are named entity recognition; terms, concepts, taxonomies and open relation extraction; coreference resolution; sentiment analysis; ontology development and alignment. The currently available resources, developed by our team over the years, related to these research topics are described in this paper.

2 Named Entity Recognition

Named Entity Recognition (NER) consists of the identification and classification of linguistic expressions identification and classification, mostly proper nouns that refer to a specific entity in the text. In general, a NER task is divided into two phases: Named Entities (NEs) identification and NEs classification. NER main challenges in the entities recognition process are the NEs delimitation during the identification phase and the ambiguity of words in the classification phase.

To deal with this task we developed the Named Entities Recognition Portuguese-Conditional Random Fields (NERP-CRF) system [1], its first version was based

on the HAREM corpus[3] and categories. More recently we are investigating geological NEs. We built a reference corpus that contains NEs considering geological classes. The corpus is formed by scientific papers and articles, thesis, and dissertations found in digital libraries. We identified eleven geological classes according to three groups: Habitat of Microfossils, Age of Rocks, and Types of Rocks. The corpus consists of 70,191 words and 3,687 geological NEs checked by a specialist, and is available at `http://www.inf.pucrs.br/linatural/NER.html`.

3 Term Extraction

Term extraction from corpora is the cornerstone of several NLP applications. A particularly interesting application of extracted terms have been developed recently to establish entity profiles [2]. An example of such profiling is available at `http://www.inf.pucrs.br/peg/lucelenelopes/profiler_PPGCC/index.html`.

Our approaches for term extraction rely on both linguistic and statistic-based techniques. The linguistic-based techniques are centered on the recognition of noun phrases from a parser annotation and a set of heuristics to increase the quality of extracted terms [3]. The statistic-based techniques intervene with the use of an index to establish the extracted term relevance, the term frequency-disjoint corpora frequency (*tf-dcf*) index [4], and, finally, with the application of cut-off policies [5]. ExATO software tool [6] implements these term extraction techniques [7]. The current version of ExATO is capable of dealing with English and Portuguese corpora in several formats of output: a concordancer, tag clouds and concept hierarchies.

To exercise our tools and techniques several domain corpora were created [8] and acquired. The corpora created are avalable at `http://www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html` and lists of the extracted terms are available at `http://www.inf.pucrs.br/peg/lucelenelopes/ll_trm.html`. Additionally, an experiment with English corpora was conducted to illustrate the impact of contrasting corpora choices in our term extraction method [9].

In [10] we proposed a method to build bilingual dictionaries for specific domains from parallel corpora. An evaluation was performed on technical manuals in English and Portuguese. The bilingual dictionaries created from the application of this method are available in `http://www.inf.pucrs.br/~linatural/multilingual/`.

4 Semantic Relation Identification

Semantic similarity could be viewed as an association of two terms, that is, the mental activation of one term when another term is presented. This idea was expressed by Zellig Harris [11] when he formulated the hypothesis that words that occur in the same contexts tend to have similar meanings. Models built on this assumption are called Distributional Similarity Models (DSMs) and take

[3] http://www.linguateca.pt/harem/

into account the co-occurrence distributions of the words in order to cluster them together [12]. In [13], we perform an evaluation on methods that use different co-occurrence orders to get similarity between terms.

In order to evaluate such methods it is also important to have datasets manually evaluated by domain experts. An important resource for evaluation in English has been defined by Rubenstein and Goodenough [14]. This dataset (RG65) contains judgements scaled from 0 to 4 according to their similarity of meaning from 51 human subjects for 65 word pairs. Following the work by Rubenstein and Goodenough, we translated into Portuguese all pairs from RG65 and evaluate them using 50 human subjects (Granada *et al.* [15]). These lists are available at `http://www.inf.pucrs.br/linatural/wikimodels/similarity.html`. Human scores are compared with previous works and an automatic evaluation is performed by comparing with models generated from Wikipedia articles.

5 Taxonomic Relations Extraction

Many methods have been proposed to extract taxonomic relations from texts. Hearst [16] proposed the extraction of taxonomic relations from texts by using lexico-syntactic patterns in the form of regular expressions, Radford [17] identifies the taxonomic relation between terms using the head of the noun phrase, since it determines the nature of the overall phrase. Using a statistical approach Caraballo [18] uses hierarchical clustering in order to identify hierarchical relations. Weeds *et al.* [19] identify relations based on their distributional inclusion, *i.e.*, two words have taxonomic relation if both share a great number of contexts. Sanderson and Croft [20] present the document subsumption method which identify taxonomic relation based on the probabilities of term co-occurrences in documents. Santus *et al.* [21] used an entropy-based measure for the unsupervised identification of taxonomic relations in DSMs.

We developed the HREx framework (`https://github.com/rogergranada/HREx`) to perform automatic and manual evaluations for relations generated by the methods presented above (Granada [22]). The framework is developed in Python and implements: (i) methods for extraction of taxonomic relations based on rules, such as patterns[16] and head-modifier[17]; and (ii) statistical methods based on hierarchical clustering[18], distributional inclusion[19], document subsumption[20] and entropy[21].

6 Open Relation Extraction

Open relation extraction systems aim at identifying all possible relations from an open-domain corpus, with no pre-specified definition of the relations [23]. These systems aim at extracting relation triples from corpus without requiring human supervision. In relation triples such as (E1, Rel, E2), E1 and E2 denote entities (represented usually by nouns or noun phrases), and Rel denotes a relation holding between E1 and E2.

In [24], we extracted relations between named entities in the Organisation domain, using Conditional Random Fields (CRF). Different feature configurations for CRF based on lexical, syntactic and semantic information have been evaluated. The evaluation was based on a subset of HAREM corpus[4] to which we added an extra annotation layer. Our annotation considered the relation descriptors ocurring between named entities of the following categories: Organisation, Person and Place. Relation descriptors are defined as the text chunks that describe an explicit relation between these entities in a sentence. For example, we have the relation descriptor *"diretor de"* (*"director of"*) that occurs between the named entities *"Ronaldo Lemos"* and *"Creative Commons"* in the sentence *"Ronaldo Lemos,* ***diretor da*** *Creative Commons, [...]"*. The annotation was performed by two linguists. Given two named entities occurring in the same sentence, if there is a text sequence (descriptor) that describes an explicit relation between these entities, it is annotated.

Based on this data, in [25], we evaluated a CRF classifier for the extraction of relation descriptors between pairs of named entities (organisations and persons - organisations and places), and also the extraction of pre-defined relation types between these entities (*"affiliation"* and *"placement"*). The resources produced in this work, texts and corresponding manually annotated triples (NE1, relation descriptor, NE2), are available at `http://www.inf.pucrs.br/linatural/data_set_RE.html`.

7 Coreference Resolution

Coreference resolution is the process of identifying mentions to the same entity in a text. In other words, this process consists of identifying the set of expressions that refer to a specific entity. For example, "The opinion is from Miguel Guerra. The agronomist...". In this case, the noun phrase "The agronomist" is coreferent with "Miguel Guerra". In [26] we propose a rule-based approach to solve coreference in Portuguese. Basically, our model is an adaptation of the system by Lee *et al.* [27], solving coreference for nominal nps, using plain texts as input. In a more recent work [28] we investigate sematic knowledge (Hyponymy and Synonymy) based on the relations provided by Onto.PT[29]. Our new semantic model is available at `http://ontolp.inf.pucrs.br/corref/`.

As part of this research we also developed Summ-it++[30], a new enriched version of the Summ-it corpus [31]. This new version adds two annotation layers to the previous coreference annotation: named entities and relations between named entities (based on the works described in Sections 2 and 6). Besides, the annotation format was changed to a well-known and widely used standard, the SemEval [32]. Summ-it++ is available at `http://www.inf.pucrs.br/linatural/summit_plus_plus.html`.

[4] http://www.linguateca.pt/harem/

8 Sentiment analysis

Sentiment analysis studies methods to analyze people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities and their aspects. This field is also known as: opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, *etc.* [33].

In [34], we proposed a lexicon-based approach to sentiment analysis in short text media, applying it to analyse Twitter messages. This approach has been later extended, in [35], to perform entity-centric sentiment analysis in Twitter messages. The process consists of identifying to which named entity in the message, each opinion-bearing expressions refers to. The reference disambiguation is achieved using a SVM machine. As part of this work we developed the opinion lexicon OpLexicon [36], available in `http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php`

In Freitas [37], we propose a sentiment analysis methodology based on features and ontologies. Initially, the method receives as input a set of reviews, which are preprocessed. After, features are identified in the preprocessed reviews using a domain ontology. The polarity is identified in the reviews considering features and using available Portuguese sentiment lexicons and linguistic rules. Finally, a summary with features and their respective polarities is generated. In [38], we analysed three different POS tagger tools to choose the best one for our experiments. We also analysed four different sentiment lexicons: SentiLex [39], Brazilian Portuguese Linguistic Inquiry and Word Count dictionary[5], synsets with polarities of Onto.PT [40] and the one we developed, OpLexicon [36].

9 Ontology development and alignment

On the area of ontology development we are studying ontology and multi-agent technologies. In this research direction, we aim to provided a tool for engineering multi-agent systems (MAS) using an ontology as a meta-model [41]. That work extends our ideas towards models of MAS represented as abstractions in ontologies [42,43]. A video that briefly demonstrates our multi-agent system engineering tool based on ontologies can be found in `https://www.youtube.com/watch?v=Lt5ZVG1cgBQ`.

We are also dealing with ontology alignment in two main fronts (i) alignment between top-level and domain ontology and (ii) ontology alignment visualization. In the first case, we are analysing the behavior of state-of-the-art matching systems to align different kinds of ontologies (domain and top-level). A top-level ontology is a high-level and domain independent ontology. The concepts expressed are intended to be basic and universal to ensure generality and expressivity for a wide range of domains [44]. Our goal is to improve the process of matching top-level and domain ontologies. In the second case, we built an environment

[5] http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

for handling ontology alignments with a visual approach: VOAR (Visual Ontology Alignment Environment) [45], available at (`http://voar.inf.pucrs.br`). Within this graphical environment, users can manually create, suppress and edit correspondences and apply a set of operations on alignments (filtering, merge, difference, etc.). Evaluation of multiple alignments, against a reference one, can be carried out with both qualitative and quantitative metrics. Finally, in its most recent version [46], VOAR allows the visualization of multiple alignments together from a set of previously loaded or manually created alignments.

10 Conclusion

In this paper we presented an overview of currently available language resources related to research in information extraction and semantic computing that we have produced at our research lab. A summary of these resources with their access links is given below.

- Named Entity Recognition
  - NE annotated corpus - geological entities: `http://www.inf.pucrs.br/linatural/NER.html`
- Term Extraction
  - Domain corpora: `http://www.inf.pucrs.br/peg/lucelenelopes/ll_crp.html`
  - List of concepts: `http://www.inf.pucrs.br/peg/lucelenelopes/ll_trm.html`
  - English-Portuguese IT dictionary and parallel corpora: `http://www.inf.pucrs.br/~linatural/multilingual`
- Semantic relation identification
  - List of semantically related pairs: `http://www.inf.pucrs.br/linatural/wikimodels/similarity.html`
- Taxonomic relations extraction
  - HREx framework: `https://github.com/rogergranada/HREx`
- Open Relation Extraction
  - Corpus and relation triples: `http://www.inf.pucrs.br/linatural/data_set_RE.html`
- Coreference resolution
  - CORP: `http://ontolp.inf.pucrs.br/corref/`
  - Summ-it++: `http://www.inf.pucrs.br/linatural/summit_plus_plus.html`
- Sentiment analysis
  - OPLexicon: `http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php`
- Ontologies
  - VOAR - alignment visualization: `http://voar.inf.pucrs.br`

We are happy to share the above research resources with the community. Our current and future efforts are related to the improvement, integration and visualization of the information provided in these resources.

Acknowledgments. This work is partially supported by CNPq, CAPES and FAPERGS.

References

1. do Amaral, D.O.F., Vieira, R.: NERP-CRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. linguamatica **6**(1) (2014) 41–49
2. Lopes, L., Vieira, R.: Building and applying profiles through term extraction. In: Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology. STIL 2015, Natal, RN, Brazil, IEEE Press (2015) 193–196
3. Lopes, L., Vieira, R.: Heuristics to improve ontology term extraction. In: PROPOR 2012 – International Conference on Computational Processing of Portuguese Language. LNCS vol. 7243 (2012) 85–92
4. Lopes, L., Fernandes, P., Vieira, R.: Estimating term domain relevance through term frequency, disjoint corpora frequency - tf-dcf. Knowledge-Based Systems **97** (2016) 237 – 249
5. Lopes, L., Vieira, R.: Evaluation of cutoff policies for term extraction. Journal of the Brazilian Computer Society **21**(1) (2015)
6. Lopes, L., Fernandes, P., Vieira, R., Fedrizzi, G.: ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora. In: Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '09), Poznan, Poland (2009) 427–431
7. Lopes, L.: Extração automática de conceitos a partir de textos em língua portuguesa. PhD thesis, PUCRS University - Computer Science Department, Porto Alegre, Brazil (2012)
8. Lopes, L., Vieira, R.: Building domain specific parsed corpora in portuguese language. In: Proceedings of the X National Meeting on Artificial and Computational Intelligence (ENIAC). (2013) 1–12
9. Lopes, L., Fernandes, P., Granada, R., Vieira, R.: The impact of contrastive corpora for term relevance measures. In: 2015 Brazilian Conference on Intelligent Systems, IEEE (2015) 146–151
10. Hilgert, L.W., Lopes, L., Freitas, A., Vieira, R., Hogetop, D.N., Vanin, A.A.: Building domain specific bilingual dictionaries. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 2772–2777
11. Harris, Z.: Distributional structure. Words **10**(23) (1954) 146–162
12. Grefenstette, G.: Explorations in automatic thesaurus discovery. Kluwer Academic Publishers Norwell (1994)
13. Granada, R., Vieira, R., Lima, V.: Evaluating co-occurrence order for automatic thesaurus construction. In: Proceedings of the IEEE 13th International Conference on Information Reuse and Integration. IRI 2012 (2012) 474–481
14. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM **8**(10) (October 1965) 627–633
15. Granada, R., dos Santos, C.T., Vieira, R.: Comparing semantic relatedness between word pairs in portuguese using wikipedia. In: Proceedings of 11th PROPOR, (PROPOR 2014). (2014) 170–175

16. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics - Volume 2, Stroudsburg, PA, USA, Association for Computational Linguistics (1992) 539–545
17. Radford, A.: Syntax: A minimalist introduction. Cambridge University Press (1997)
18. Caraballo, S.: Automatic construction of a hypernym-labeled noun hierarchy from text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. (1999) 120–126
19. Weeds, J., Weir, D., McCarthy, D.: Characterising measures of lexical distributional similarity. In: Proceedings of the 20th International Conference on Computational Linguistics. COLING-2004 (2004) 1015–1021
20. Sanderson, M., Croft, B.: Deriving concept hierarchies from text. In: Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval. (1999) 206–213
21. Santus, E., Lenci, A., Lu, Q., Schulte im Walde, S.: Chasing hypernyms in vector spaces with entropy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2014) 38–42
22. Granada, R.: Evaluation of methods for taxonomic relation extraction from text. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul – Université Toulouse III – Paul Sabatier (2015)
23. Banko, M., Etzioni, O.: The tradeoffs between open and traditional relation extraction. In McKeown, K., Moore, J.D., Teufel, S., Allan, J., Furui, S., eds.: Proceedings of ACL-08: HLT, Columbus, Ohio, Association for Computational Linguistics (2008) 28–36
24. Collovini, S., Pugens, L., Vanin, A.A., Vieira, R.: Extraction of relation descriptors for portuguese using conditional random fields. In: In Proceedings of Advances in Artificial Intelligence - IBERAMIA 2014 - 14th Ibero-American Conference on Artificial Intelligence, Santiago de Chile, Chile (2014) 108–119
25. Collovini, S., Machado, G., Vieira, R.: A sequence model approach to relation extraction in portuguese. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016). (In Press)
26. Fonseca, E.B., Vieira, R., Vanin, A.: Adapting an entity centric model for portuguese coreference resolution. In: Proceedings of the 10th Annual Conference on Language Resources and Evaluation (LREC 2016), In Press (2016)
27. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics **39**(4) (2013) 885–916
28. Fonseca, E.B., Vieira, R., Vanin, A.: Improving coreference resolution with semantic knowledge. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR 2016), In Press (2016)
29. Oliveira, H.G., Gomes, P.: Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. Language Resources and Evaluation **48**(2) (2014) 373–393
30. Antonitsch, A., Figueira, A., Amaral, D., Fonseca, E., Vieira, R., Collovini, S.: Summ-it++: an enriched version of the summ-it corpus. In: Proceedings of 10th edition of the Language Resources and Evaluation Conference (LREC 2016), In Press (2016)
31. Collovini, S., Carbonel, T.I., Fuchs, J.T., Coelho, J.C., Rino, L., Vieira, R.: Summ-it: Um corpus anotado com informaes discursivas visando a sumarizao automática.

In: Proceedings of V Workshop em Tecnologia da Informao e da Linguagem Humana. (2007) 1605–1614
32. Recasens, M., Màrquez, L., Sapena, E., Martí, M.A., Taulé, M., Hoste, V., Poesio, M., Versley, Y.: Semeval-2010 task 1: Coreference resolution in multiple languages. In: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics (2010) 1–8
33. Liu, B.: Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). California: Morgan & Claypool Publishers (2012)
34. Souza, M., Vieira, R.: Sentiment analysis on twitter data for portuguese language. In: Computational Processing of the Portuguese Language. Springer Berlin Heidelberg (2012) 241–247
35. Souza, M., Vieira, R.: Entity-centric sentiment analysis on twitter data for the potuguese language. In: 9th Brazilian Symposium in Information and Human Language Technology (STIL 2013), Fortaleza, Brazil. (2013)
36. Souza, M., Vieira, R., Chishman, R., Alves, I.M.: Construction of a portuguese opinion lexicon from multiple resources. In: 8th Brazilian Symposium in Information and Human Language Technology. (2011) 59–66
37. Freitas, L.: Feature-level sentiment analysis applied to brazilian portuguese reviews. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul (2015)
38. Freitas, L., Vieira, R.: Comparing portuguese opinion lexicons in feature-based sentiment analysis. International Journal of Computational Linguistics and Applications **1**(4) (2013) 147–158
39. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: 10th International Conference Computational Processing of the Portuguese Language. (2012) 218–228
40. Oliveira, H.G., Santos, A.P., Gomes, P.: Assigning polarity automatically to the synsets of wordnet-like resource. In: 3rd Symposium on Languages, Applications and Technologies. (2014) 169–184
41. Freitas, A., Hilgert, L., Marczak, S., Meneguzzi, F., Bordini, R.H., Vieira, R.: A multi-agent systems engineering tool based on ontologies. In: 34th International Conference on Conceptual Modeling, Stockholm, Sweden. Lecture Notes in Computer Science, Springer (2015)
42. Freitas, A., Bordini, R.H., Meneguzzi, F., Vieira, R.: Towards integrating ontologies in multi-agent programming platforms. In: 2013 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2013, Atlanta, Georgia, USA. (2013)
43. Freitas, A., Schmidt, D., Panisson, A., Meneguzzi, F., Vieira, R., Bordini, R.H.: Applying ontologies and agent technologies to generate ambient intelligence applications. In: Joint Proceedings Collaborative Agents – Research & Development, CARE for Intelligent Mobile Services & Agents, Virtual Societies and Analytics. (2014) 22–33
44. Semy, S.K., Pulvermacher, M.K., Obrst, L.J.: Toward the use of an upper ontology for u.s. government and u.s. military domains: An evaluation. Technical report, Submission to Workshop on Information Integration on the Web (IIWeb-04) (2004)
45. Severo, B., Trojahn, C., Vieira, R.: VOAR: A visual and integrated ontology alignment environment. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014. (2014) 3671–3677
46. Severo, B., Trojahn, C., Vieira, R.: A gui for visualising and manipulating multiple ontology alignments. In: International Semantic Web Conference (Posters & Demos). Volume 1486. (2015)

MWE-aware corpus processing with the mwetoolkit and word embeddings

Silvio Cordeiro[1,2], Carlos Ramisch[2], Marco Idiart[3], and Aline Villavicencio[1]

[1] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[2] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
srcordeiro@inf.ufrgs.br,carlos.ramisch@lif.univ-mrs.fr,
marco.idiart@gmail.com,avillavicencio@inf.ufrgs.br

Abstract. Multiword expressions (MWEs) are an integral part of language whose importance has long been recognized. However, their heterogeneous characteristics have proved a challenge to computational tasks and applications, including machine translation. In this paper we discuss how MWEs can be dealt with using the mwetoolkit, a language-independent platform for MWE related tasks. In particular, we concentrate on 3 tasks: (1) corpus processing for type identification from corpora, (2) token identification and corpus annotation, and (3) the construction of semantic distributional models for compositionality detection based on word embeddings. The mwetoolkit provides a uniform platform for creating MWE resources, and we discuss its use for both English and Portuguese MWE processing.

Keywords: Multiword expressions, token identification, compositionality detection.

1 Introduction

For many natural language processing tasks, such as machine translation and text simplification, improvements in quality require precise treatments not only for single words, but for sequences of words that act as a unit at some level of linguistic analysis [4], known as multiword expressions (MWEs). These include semantic units that often span over multiple lexemes in the text and whose precise interpretation may not be straightforwardly inferred from the individual meaning of their component words [5], like lua de mel (*honeymoon*), faz de conta (*make believe*), chover no molhado (*preach to the converted*), grosso modo (*roughly*). Beyond semantic non-compositionality, MWEs may also present some lexical, syntactic, pragmatic, statistical and/or other semantic idiosyncrasies [2]. As a consequence, if determining the meaning of single words is a difficult task on its own, dealing with MWEs often requires also determining how the words affect one another in possibly unpredictable ways. In short, NLP technologies that involve some level of semantic processing need to take MWEs into account to obtain accurate results.

With the growing interest in processing MWEs, there is an increasing need for resources that represent their linguistic and distributional characteristics and for tools that enable the construction of these resources. This often involves supporting tasks like lexicon construction, through type discovery in corpora; corpus annotation, by means of token identification of MWE entries in corpora; and (distributional) semantic model construction from MWE-annotated sentences. All of these tasks can be done in a single platform: the mwetoolkit [24], which is a language-independent framework that has been successfully used for modeling lexical, syntactic and semantic characteristics of MWEs in many languages. For instance, it offers a variety of association scores that estimate the degree of statistical interdependence of words based on corpus frequencies. It can also be applied on distributional semantic models (DSMs) to determine semantic compositionality of MWEs. Finally, building MWE-aware resources requires careful corpus processing, since MWEs are often not dealt with adequately by POS taggers and parsers for these languages. The mwetoolkit provides user-customizable flexible search capabilities for annotating known MWEs in corpora.

In this paper, we discuss the use of the mwetoolkit for generic corpus processing, MWE token-level identification and compositionality prediction, concentrating on English and Portuguese MWEs. The mwetoolkit allows treating idiomatic expressions as semantic units and representing compositional expressions as the combination of individual meanings. This information can in turn be exploited by NLP systems in tasks such as machine translation. The framework is freely available as part of the mwetoolkit distribution[1]. This paper is structured as follows: in §2 we discuss some techniques for MWE treatment. In §3 we introduce the mwetoolkit and its application for MWE tasks in §4, §5 and §6. We finish with conclusions and future work.

2 Related Work

Traditional corpus processing pipelines deal with incremental abstraction levels, going from tokens to words, parse trees, lexical senses and argument structure. However, they often ignore the fact that MWEs are frequent in most languages and domains, lacking precision and linguistically accurate representations as far as multiword phenomena are concerned.

MWE-aware corpus processing includes the discovery of new multiword lexical units in corpora, token-based identification in running text, and compositionality prediction. As a basis for MWE discovery, linguistic knowledge has been used to target specific kinds of MWE, like compound nouns or phrasal verbs, often in combination with frequency information and association measures, which offer an inexpensive language and type independent means of detecting recurrent patterns [17, 16, 12, 32, 24]. The mwetoolkit in particular includes support not only for MWE discovery, but also for token identification and for the use of distributional semantic models in compositionality prediction.

[1] `http://mwetoolkit.sf.net`

The accurate identification of MWE tokens in running text is a fundamental task in the pipeline of many NLP applications. For example, MT systems need to know when a group of words must be translated as a unit, and parsers need to recognize the cases where a seemingly unrelated set of words should be grouped as a single lexeme or constituent. A toolkit such as `jMWE` [18] can be used to annotate sentences based on preexisting lexicons. The output is a corpus where each MWE occurrence found in a lexicon has been matched and tagged. Finite-state transducers can also be used to take into account the internal morphology of component words and perform efficient tokenization based on MWE dictionaries [1, 30]. The problem of MWE identification has also been modeled using supervised machine learning. Probabilistic MWE taggers usually encode the data using a begin-inside-outside scheme and learn CRF-like taggers on it [6, 31]. However, if the words in the MWE do not appear contiguously (e.g. due to internally inserted modifiers), a contiguous annotator such as `jMWE` will fail to detect them. Additionally, the use of separate tools for discovery and token identification will miss the opportunity of sharing information. The mwetoolkit addresses the latter by allowing the integration of type and toke identification, and the former by enabling non-contiguous matches (e.g. *eat FOOD up* as in *eat that wonderful chocolate cake up*) and optional and variable elements in MWEs (e.g. *throw PERSON to the lions/wolves*).

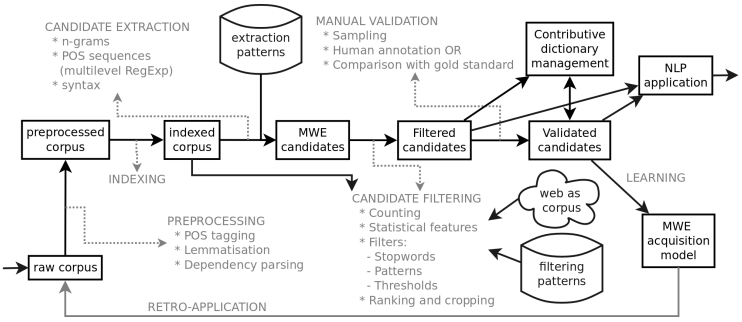


Fig. 1. Overview of the architecture of the mwetoolkit.

In terms of semantics, MWEs can range from compositional cases, like *campeonato de tênis (tennis championship)*, to fully idiomatic or non-compositional cases, like *bola nas costas* (lit. *ball on the back*, meaning *betrayal*). Thus, MWE compositionality detection, requires techniques for calculating their degree of compositionality. A strategy that is often employed is to compare characteristics of an MWE with those of its individual components, usually using information from manually constructed resources such as WordNet [14]or from automatically constructed DSMs [15]. Since the former are not always available for a given language or domain, we concentrate on the latter, given that many tools

are available for building them, like Dissect [9], minimantics[2], word2vec[3] [21] and Glove[4] [23]. As word meanings can be represented as vectors, composition can be effectively modeled through simple operations like vector addition and multiplication [22]. For noun compounds, Reddy et al. [28] suggest a compositionality measure which is the cosine similarity between the MWE vector and the sum of the vectors of the component words. This model was also used by Salehi et al. [29], in combination with word translation information coming from parallel corpora, while Yazdani et al. [33] propose more sophisticated composition functions, based on linear, non-linear and neural network projections. Compositionality detection in the mwetoolkit is based on vector addition and cosine similarity.

3 Corpus processing

Most of our corpus processing is based on the mwetoolkit as it is a collection of general tools for language-independent corpus and lexicon preprocessing. Though it was initially conceived with the goal of MWE discovery and extraction, the toolkit has grown to encompass many other functions, including single-word-level operations, in-corpus MWE identification (see Section 5), evaluation of corpus annotations, and transformation among corpus formats. This latter is particularly useful: as the mwetoolkit supports multiple file formats (PALAVRAS, TreeTagger, CONLL, RASP...)[5],one is not constrained to working with the output format of a particular tool. Besides using it to convert between file formats, we also use the toolkit for correcting tokenization problems, case homogenization, and transformation between tagsets. The tools that are provided in the mwetoolkit can be chained as a sequence of operations in a pipeline, or just be used in isolation (for example, to extract from a corpus a list of MWE candidates along with their association measures). The general architecture of the tool is presented in Figure 1. Although not particularly interesting from a research point of view, it is important to have straightforward procedures and tools, as generic as possible, for efficient and reproducible corpus-based NLP.

The underlying architecture of the toolkit also allows for an easy integration of new input/output formats, which can also be developed by the toolkit users themselves. Besides corpora, there is native support for other kinds of information, such as MWE candidate lists and word embeddings. This allows the toolkit to provide operations such as pattern-based MWE annotation and filtering subsets of word embeddings.

[2] `https://github.com/ceramisch/minimantics`
[3] `http://word2vec.googlecode.com/svn/trunk`
[4] `https://github.com/stanfordnlp/GloVe`
[5] See `http://mwetoolkit.sf.net` for file format details.

4 MWE Type Discovery

The mwetoolkit uses an extraction algorithm that builds on the notion of regular-expression patterns based on token properties. For example, given a noun compound pattern such as `Noun Noun`$^+$ and a POS-tagged corpus, the extractor lists all occurrences of this expression in the text. The automatic discovery of new MWE lexical types is performed in two steps. The first step, *candidate generation*, can be done using a combination of flat linguistic information from surface forms, lemmas and parts of speech (POS) tags (e.g. `VERB NOUN` and *take* `NOUN`) and even include wildcards that stand for any word or POS. In the second step, *candidate filtering*, a set of association measures are calculated for each candidate to filter as much noise as possible among the candidates.[6]. Additionally, if a gold standard is available then the toolkit can further provide annotation to build a classifier, automatically annotating each candidate to indicate whether it is contained in the gold standard or not.

Our long-term goal is to create and enrich lexicons of MWEs for under-resourced languages (in terms of MWEs) like Portuguese. Such resources will then be integrated with more traditional NLP tools like POS taggers and parsers, probably involving some pre-identification of potential MWE units, as described in Section 5.

5 MWE Token Identification

MWE token identification can be seen as a tagging process that takes as input a corpus and, optionally, an MWE lexicon, outputting an annotated corpus that explicitly indicates where each expression occurs. This indication can range from simply joining the MWE components as a single word (using a special "MWE separator") to more complex metadata representations (e.g. for corpora represented in XML, one may indicate each MWE by its word indexes). Token identification capabilities in the mwetoolkit include:

1. Different gapping possibilities
   - Contiguous: Matches contiguous sequences of words from a list of MWEs.
   - Gappy: Matches words with up to a limit number of gaps in between.
2. Different match distances
   - Shortest: Matches the shortest possible candidate (e.g. for phrasal verbs, we want to find only the closest particle).
   - Longest: Matches the longest possible candidate (e.g. for noun compounds).
   - All: Matches all possible candidates (useful as a fallback when shortest and longest are too strict).
3. Different match modes
   - Non-overlapping: Matches at most one MWE per word in the corpus.

[6] The detailed description of the four association measures generated by the toolkit can be found in [24], and a comparison with other tools is presented in [26]

  - Overlapping: Allows words to be part of more than one MWE (e.g. to find MWEs inside the gap of another MWE).
4. Source-based annotation: Uses the information retrieved in automatic MWE type discovery (see Section 4). Since MWEs are extracted with detailed information about the source corpus and sentence, this can later be used for quick annotation of the original corpus.

Given an input such as Figure 2.1 and the two MWE patterns described by the POS regular expressions below, the gappy approach with different match distances will detect different types of MWEs: using the *longest* match distance (Figure 2.2), the *shortest* match distance (Figure 2.3) and one per MWE type (Figure 2.4).

- `NounCompound` → `Noun Noun`$^{+}$
- `PhrasalVerb` → `Verb (Word`*`) Particle`

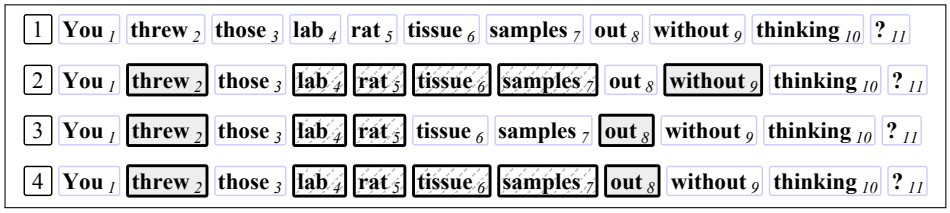


Fig. 2. Gappy MWE-annotated output with different match distances.

The toolkit enables both the annotation of a corpus based on a preexisting lexicon of MWEs or the combination of type-based discovery and corpus annotation, first generating a lexicon which is subsequently used for annotating the corpus. When annotating the same corpus from which MWE types were extracted, source-based annotation can be used for best results.

6 Semantics and Compositionality

High-quality semantic resources such as WordNet are not available for most languages and domains, and often they do not include many MWEs. Therefore, we perform semantic processing using distributional semantic models (DSMs) built from large unannotated corpora. Given the many state-of-the-art tools available for building DSMs, we assume that word embeddings are built offline by one of the available dedicated tools. We have augmented the mwetoolkit with internal file readers that enable the automatic detection and reading of a variety of embedding formats, including the *Minimantics*, *word2vec* and *GloVe* formats.

File readers see the word embedding files as a list of named embeddings, each of which associates a target word form (e.g. its lemma) to a mapping between context identifiers and real values. For fixed-length embeddings, where there are no clear semantics attached to each dimension, we read the file as if each of the n

values corresponded to artificial context identifiers $[c_0, \ldots, c_{n-1}]$. On the other hand, in models such as Minimantics, the context identifiers are the context words themselves, as they appear in text.

For the semantic processing, the toolkit provides `feat_compositionality` [7], which outputs a compositionality score for each MWE in a list of input candidates, based on an input word-embeddings file.[7] In the first step, it combines the vectors $\overrightarrow{w_i}$ representing each word w_i in an MWE, using one of the available operators:

- *PointwiseAddition*: where pointwise vector addition is used to combine two or more embeddings [21]. Weights can be applied by explicitly specifying a list of multiplicative constants α_i, one for each component word w_i.
- *PointwiseMultiplication*: using pointwise vector multiplication, where each element of one vector is scaled by the respective elements in the other words of the MWE [22].

Once the embeddings of the words inside the MWE candidate have been combined (e.g. *bounty*⊕*hunter*), it is possible to compare the result with the embedding of the MWE itself (in this case, the embedding for the token *bounty_hunter*) using cosine similarity. The compositionality score using weighted pointwise addition for an MWE candidate composed of words w_1 through w_m is:[8]

$$comp(w_1 \ldots w_m) = \cos\left(\frac{\overrightarrow{w_1 \ldots w_m}}{||\overrightarrow{w_1 \ldots w_m}||}, \sum_{j=1}^{m} \alpha_j \frac{\overrightarrow{w_j}}{||\overrightarrow{w_j}||}\right)$$

7 Evaluation

The mwetoolkit has been used for Portuguese type identification for MWEs in general [20] and for specific MWE types, like support verb constructions [10, 11] and noun–adjective compounds [25]. For English, it has also been used for MWEs in general [26, 8], and for a variety of specific MWE types such as verb–particle constructions [27] and compound nouns [7, 25]. For instance, in an evaluation of these identification methods, we have obtained an F-score of 51.48% for MWE identification using both pattern-based and training-based methods [8]. In this same evaluation, we found that compound nouns account for the greatest proportion of MWEs in both the training and the testing corpora. The domain of the corpus did not seem to have a great influence on our method's performance,

[7] For imputation of missing values the mwetoolkit allows two strategies: in the first, if a single-word is not found in the embeddings file, the zero vector is used; while in the second, if the MWE itself is missing in the corpus, the average compositionality score of all other MWEs in the list of candidates is used, as per Salehi et al. [29].

[8] Even though normalization is not strictly necessary, it does not influence the results of cosine similarity and would allow the generalization to similarity measures other than cosine.

even though the corpora ranged from informal written texts (twitter) to formal presentations (TED talks).

Moreover, as the toolkit is language independent, it also facilitates multilingual work. For instance, it was used to construct a multilingual dataset of human compositionality judgments for compound nouns in Portuguese, English and French [25]. This work required type identification for the selection of compound nouns to be evaluated in each language, given a frequency threshold that aims to obtain only familiar compounds. It also involved token identification for selecting sentences to serve as context for the human judgments. Finally, the annotated corpus was used to construct DSMs and calculate compositionality scores.

As an example of task that requires intensive MWE-aware corpus processing, we describe in more details an evaluation of compositionality prediction for compound nouns in English.

7.1 Compositionality

To evaluate compositionality prediction, we use a set of 1042 English noun compounds [13]. We train an instance of each of these distributional semantic models: *Minimantics*, *word2vec (cbow)* and *GloVe*. For training, we feed an MWE-annotated corpus where MWEs are joined as a single token. We fix the following parameters:

- Corpus: UKWaC, containing 2G words of English texts crawled from the web [3];
- Context window: lemma of each content word 8 words to the left/right of the target;
- Context weight decay: linear, that is, $\left[\frac{8}{8}, \frac{7}{8}, \frac{6}{8}, \ldots, \frac{1}{8}\right]$ [19];
- Dimensions per embedding: 250.

We compare our model for compositionality prediction with a simple baseline that uses the *log-likelihood* (LL) association score. We implemented several evaluation measures used in the literature to compare the model predictions with human judgments, among them Spearman's Rho (ρ), Normalized Discounted Cumulative Gain (NDCG), Best F-score (F_1), Precision at k (P@k) and Average precision (AP). Each compound has four binary judgments, and for ρ, we use the sum of the binary judgments to rank the compounds, while for NDCG, F_1, P@k and AP, a compound is considered compositional if at least two judges consider it so, following the heuristic adopted by Yazdani et al. [33] to allow a comparison with their results. Table 1 presents the results,[9] where all measures range from 0 to 1 (except for ρ, which ranges from -1 to 1); values close to 1 indicate better results. The prediction based on distributional models correlates much better than the baseline with the human non-compositionality scores. These results are comparable with what has been found in other works [33], even though we have not tuned the parameters of our models.

[9] Except for the baseline, all of the other models use a 50% : 50% combination weight (i.e. an average between the vector of the head and the vector of the modifier).

	ρ	NDCG	F_1	P@100	AP
Baseline (LL)	-0.19	0.63	0.32	0.09	0.15
Minimantics	0.17	0.72	0.36	0.32	0.27
word2vec	**0.31**	**0.84**	**0.46**	**0.46**	**0.40**
GloVe	0.07	0.68	0.35	0.14	0.21
Yazdani2015	0.41	0.86	0.49	0.54	N/A

Table 1. Evaluating for non-compositionality

8 Conclusions and Future Work

In this paper we discussed some of the practical issues involved in MWE-aware corpus processing, from type and token identification to compositionality detection. We describe how these issues are addressed by the mwetoolkit, which is an integrated framework for processing MWEs. The toolkit offers semantic and word embedding capabilities in addition to the standard techniques for lexical and syntactic MWE representation. The results we obtained in the evaluation are compatible with the state of the art even with simple methods and without any optimization. The mwetoolkit was used to create a variety of resources for both Portuguese and English, ranging from MWE lists and annotated corpora, to resources containing MWE compositionality judgments. All these resources are needed for high-precision NLP tasks and applications, including Machine Translation and Text Simplification.

As future work, we plan on providing support for other sense composition functions (e.g. matrices, tensors and neural networks). We also intend on performing an extensive evaluation of techniques examining MWE compositionality on other datasets and languages. Additionally we envisage applying the same methodology to corpora abiding by the Portuguese Language Orthographic Agreement, to assess how the agreement might have had an impact on the identification of MWEs. Our goal in the future is to use this information in MT systems, to better translate MWEs, which are currently a great challenge for MT technology.

9 Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche through projects PARSEME-FR (ANR-14-CERA-0001) and ORFEO (ANR-12-CORP-0005), and by French-Brazilian cooperation projects CAMELEON (CAPES-COFECUB #707/11) and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7). Part of the results presented in this paper were obtained through projects "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91, and CNPq (482520/2012-4 and 312114/2015-0).

References

1. Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A.: Open-source portuguese–spanish machine translation. In: Computational Processing of the Portuguese Language, pp. 50–59. Springer (2006)
2. Baldwin, T., Kim, S.N.: Multiword expressions. In: Handbook of Natural Language Processing, Second Edition., pp. 267–292 (2010)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation 43(3), 209–226 (2009), http://www.springerlink.com/content/C348PU7321GX5081
4. Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (May 2002), http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf, ACL Anthology Identifier: L02-1259
5. Choueka, Y.: Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In: Fluhr, C., Walker, D.E. (eds.) Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988). pp. 609–624. CID, Cambridge, MA, USA (Mar 1988)
6. Constant, M., Sigogne, A.: MWU-aware part-of-speech tagging with a CRF model and lexical resources. In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World. pp. 49–56. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/W11-0809
7. Cordeiro, S.R., Ramisch, C., Villavicencio, A.: mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA) (2016), to appear
8. Cordeiro, S.R., Ramisch, C., Villavicencio, A.: UFRGS&LIF: Rule-based MWE identification and predominant-supersense tagging. In: International Workshop on Semantic Evaluation (Sem-Eval 2016). Association for Computational Linguistics (2016), to appear
9. Dinu, G., Pham, N.T., Baroni, M.: Dissect - distributional semantics composition toolkit. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 31–36. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/P13-4006
10. Duran, M.S., Ramisch, C.: How do you feel? Investigating lexical-syntactic patterns in sentiment expression. In: Proceedings of Corpus Linguistics 2011: Discourse and Corpus Linguistics Conference. Birmingham, UK (Jul 2011)
11. Duran, M.S., Ramisch, C., Aluísio, S.M., Villavicencio, A.: Identifying and analyzing Brazilian Portuguese complex predicates. In: Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World. pp. 74–82. MWE '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)

12. Evert, S., Krenn, B.: Using small random samples for the manual evaluation of statistical association measures. Computer Speech & Language 19(4), 450–466 (2005)
13. Farahmand, M., Smith, A., Nivre, J.: A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In: Proceedings of the 11th Workshop on Multiword Expressions. pp. 29–33. Association for Computational Linguistics, Denver, Colorado (June 2015), http://www.aclweb.org/anthology/W15-0904
14. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). MITPRESS (May 1998), 423 p.
15. Ferret, O.: Compounds and distributional thesauri. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 2979–2984. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/754_Paper.pdf, ACL Anthology Identifier: L14-1590
16. Frantzi, K., Ananiadou, S., Mima, H.: International Journal on Digital Libraries 3(2), 115–130 (2000)
17. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering 1(01), 9–27 (1995)
18. Kulkarni, N., Finlayson, M.: jMWE: A Java toolkit for detecting multi-word expressions. In: Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World. pp. 122–124. MWE '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
19. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3, 211–225 (2015), https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570
20. de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., Villavicencio, A.: Alignment-based extraction of multiword expressions. Language resources and evaluation 44(1-2), 59–77 (2010), http://www.springerlink.com/content/H7313427H78865MG
21. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
22. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive science 34(8), 1388–1429 (2010)
23. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1162
24. Ramisch, C.: Multiword Expressions Acquisition - A Generic and Open Framework. Theory and Applications of Natural Language Processing, Springer (2015), http://dx.doi.org/10.1007/978-3-319-09207-2
25. Ramisch, C., Cordeiro, S.R., Zilio, L., Idiart, M., Villavicencio, A.: How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics (2016), to appear

26. Ramisch, C., De Araujo, V., Villavicencio, A.: A broad evaluation of techniques for automatic acquisition of multiword expressions. In: Proceedings of ACL 2012 Student Research Workshop. pp. 1–6. ACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
27. Ramisch, C., Villavicencio, A., Moura, L., Idiart, M.: Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning. pp. 49–56. Association for Computational Linguistics (2008), http://www.aclweb.org/anthology/W08-2107
28. Reddy, S., McCarthy, D., Manandhar, S.: An empirical study on compositionality in compound nouns. In: Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011). Chiang Mai, Thailand (November 2011), http://sivareddy.in/papers/ijcnlp2011empirical.pdf
29. Salehi, B., Cook, P., Baldwin, T.: A word embedding approach to predicting the compositionality of multiword expressions. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 977–983. Association for Computational Linguistics, Denver, Colorado (May–June 2015), http://www.aclweb.org/anthology/N15-1099
30. Savary, A.: Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In: Maneth, S. (ed.) CIAA. Lecture Notes in Computer Science, vol. 5642, pp. 237–240. Springer (2009), http://dblp.uni-trier.de/db/conf/wia/ciaa2009.html#Savary09
31. Schneider, N., Danchik, E., Dyer, C., Smith, N.A.: Discriminative lexical semantic segmentation with gaps: running the MWE gamut. Transactions of the Association for Computational Linguistics 2, 193–206 (Apr 2014), http://www.transacl.org/wp-content/uploads/2014/04/51.pdf
32. Seretan, V.: Syntax-Based Collocation Extraction, Text, Speech and Language Technology, vol. 44. SPRINGER, Dordrecht, Netherlands, 1st edn. (2011), 212 p.
33. Yazdani, M., Farahmand, M., Henderson, J.: Learning semantic composition to detect non-compositionality of multiword expressions. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1733–1742. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/D15-1201

ZAC: Zero Anaphora Corpus
A Corpus for Zero Anaphora Resolution in Portuguese

Jorge Baptista[1,3], Simone Pereira[1,3], and Nuno Mamede[2,3]

[1] Universidade do Algarve, Faculdade de Ciências Humanas e Sociais
Campus de Gambelas, 8005-139 Faro, Portugal
jbaptis@ualg.pt
[2] Universidade de Lisboa, Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[3] INESC-ID Lisboa/L2F – Spoken Language Lab
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
Nuno.Mamede@tecnico.ulisboa.pt

Abstract This paper describes a *corpus* of Brazilian Portuguese texts built in view of the construction of an Anaphora Resolution system, which is part of a fully-fledged Natural Language Processing system (STRING). The ZAC corpus is aimed at the resolution of the so-called *zero-anaphora*, that is, an anaphora relation where the anaphoric expression (or *anaphor*) has been zeroed The paper briefly discusses the linguistic issues in the process of zero anaphora resolution, and describes the annotation process in detail, as well as the main aspects of the anaphoric relations thus annotated.

Keywords Zero anaphora, Corpus, Brazilian Portuguese, Anaphora Resolution, Natural Language Processing

1 Introduction

The Natural Language Processing (NLP) task of Anaphora Resolution (AR) is critical for other NLP tasks, for example, for parsing and semantic role labelling, as well as for many applications such as machine translation, information extraction and question answering [1]. *Anaphora* is a major discursive device used to avoid repetition and increase the cohesion of the text, making the interpretation of a given sentence to depend upon the interpretation of previous elements [2, pp. 1701–12]. For example, in sentence (1), the (clitic) pronoun (the *anaphor*) *–la* refers to (is the *antecedent* of) the proper name *Amazônia* 'Amazonia' that appears in a previous moment in the discourse (thus, the relation between them being called *anaphora*). Besides contributing to the cohesion of the discourse, this anaphoric expression is also *co-referential*, *i.e.* both the name and the pronoun refer to the same extralinguistic entity, a geographic region in the real world:

(1) *Para salvar a* ***Amazônia*** *é preciso conhecê-****la****.* 'To save the *Amazonia* it necessary to know *her*'

Anaphora can also be classified according to the relative location of the antecedent and the anaphor: (a) *intrasentential anaphora*, if the antecedent is in the same sentence as the anaphor; (b) *intersentential anaphora*, if the anaphora relation is established across sentence boundaries; (c) *anaphora* proper, when the antecedent appears before the anaphor in the linear order of discourse; (d) *cataphora*, if that order is reversed. In addition to the immense amount knowledge that may be needed to perform anaphora resolution, the various forms that anaphora can assume make it a very challenging task, especially when one intends to "teach" computers how to solve anaphora. Machine-learning approaches to anaphora resolution, which constitute the main trend in current AR research, require large quantities of annotated corpora, where anaphoric relations are explicitly marked. Most of the previous work addressed pronominal anaphora, where the anaphor is a pronoun, as in (1). However, little work has been devoted to zero anaphora resolution and, to our knowledge, no corpus marked up with deleted subject noun phrases (NP) is available for Portuguese.

This paper presents the process of building a corpus with manually annotated zero-anaphoric relations in view of building an Zero Anaphora Resolution module [3] integrated in a fully-fledged Natural Language Processing system, STRING [4]. The paper describes the annotation process in detail, as well as the main linguistic aspects of the anaphoric relations thus annotated.

This paper is structured as follows: In the next section, we provide a brief overview of the major NLP approaches to anaphora resolution and current systems developed for Portuguese. Next, we present the corpus contents and then we describe in detail the main issues concerning zero anaphora, and the way they were annotated in the corpus. From this, the major results of the annotation process are presented. The paper concludes with some final remarks and perspectives for future work.

2 State of the Art

AR algorithms can be broadly classed into rule-based and machine learning approaches. Initially, it was the rule-based approaches such as Hobbs's algorithm [5] and Lappin and Leass's [6] resolution of anaphora procedure (RAP), which gained popularity. In the 1990s and 2000s, as people grew aware of the complexity of the job at hand, research started to be limited to specific types of anaphora in view of ultimately achieving better results. Dagan and Itai's collocation pattern-based approach [7]; Kennedy and Boguraev's parse-free approach [8]; Paraboni and Lima's research on Portuguese possessive pronominal anaphora [9]; Mitkov's algorithm [1] and Chaves and Rino's adaptation of Mitkov's algorithm for anaphora resolution in (Brazilian) Portuguese [10]; all these approaches brought new insights about AR and new ways to approach the task. Machine learning approaches to pronoun (and, in general, to anaphora and coreference) resolution [11, 12, 13, 14] have been an important direction of research. A corpus similar to ZAC has been presented for Spanish [15] but in a different theoretical framework.

In 2010, Pereira [3] presented a rule-based module for zero-anaphora resolution integrated in the STRING system [4]. The author reported a precision of 60.1%, a recall of 45.5% and a F-measure of 51.8%. In 2011, Nobre [16] implemented ARM 1.0, an adaptation of the Mitkov's algorithm for resolving Portuguese pronominal anaphora, achieving 33.5% F-measure, a value too low, compared to other state-of-the-art systems. Later, Marques [17] developed ARM 2.0 , an entirely new system, based on a hybrid, statistical and rule-based, approach, with a larger corpus, using a more complex annotation scheme [18]. The system reports 54.4% F-measure. It should be noted, however, that unlike previous work [1], no pre-processing of the corpus has been made, which renders the AR scenario more realistic. Both works [17, 16] only targeted pronominal anaphora, though.

3 Corpus Annotation

The Zero Anaphora Corpus (ZAC)[4] consists on a set of full and partial texts retrieved from the web, or digitalised from books, encompassing several genres and text types, namely journalistic and literary text from contemporary Brazilian Portuguese native-speaking authors, totalling 35,212 words. This corpus was split into two parts: the training corpus with 22,385 words (63.5%) and the evaluation corpus with 12,827 words (36.5%). Table 1 shows the breakdown per text type of the ZAC corpus current content.

Table 1. Breakdown of the contents of the ZAC corpus per text type.

	Training corpus		Evaluation Corpus		Full ZAC corpus	
Text type	**Words**	**%**	**Words**	**%**	**Words**	**%**
Special report	10,272	46	5,519	43	15,791	45
News	905	4	864	7	1,769	5
Chronicle	5,416	24	2,969	23	8,385	24
Fiction (short stories)	2,029	9	1,198	9	3,227	9
Fiction (novel)	3,763	17	2,277	19	6,040	17
Total	**22,385**		**12,827**		**35,212**	

The corpus was jointly annotated by two linguists, who revised and discussed each other's work, so that each annotation one of them encoded was always checked by the other annotator. Because of this methodology, no inter-annotator agreement measure can be provided. A set of very detailed annotation guidelines [19] were produced to help the annotation process and render it more consistent. For lack of space, only an outline of these guidelines is presented here.

The annotation of zero anaphora consisted, basically, in inserting a tag for the zero *anaphor* with the form '`[0=<x]`' in the empty slot of the zeroed constituent, linking it to its immediate *antecedent* (`x`) and determining whether it

[4] https://string.l2f.inesc-id.pt/w/index.php/Corpora [last access: 31-05-2016].

appeared *before* '<' (anaphora proper) or *after* '>' the anaphor (cataphora). Inter-sentential anaphora is marked with double arrows '<<' and '>>', irrespective of the number of intervening sentences.

Briefly, the following linguistic situations were encoded. In coordinated clauses, only the subject of explicit verbs under coordination are marked. Clausal antecendents are indicated by their main verb (5).

(5) *"**Esconder** um programa desta magnitude não é apenas inapropriado, mas* `[0(clause)=<esconder]` *é também ilegal", disse o senador democrata Dick Durbin.* 'Hiding a program of this magnitude is not only inappropriate but [it] is also illegal, said democratic senator Dick Durbin'

On coordinated relative clauses, where the second subject relative pronoun has been zeroed, this should be marked but with the special notation `[0(que)=<X]`, where `X` represents the antecedent of the zeroed relative pronoun, as seen in (6):

(6) *Os processos epigenéticos também podem ocorrer pela modificação das histonas, as **linhas** que envolvem o DNA e* [0(que)=<linhas] *formam um novelo.* 'The epigenetic process can also occur by the modification of histones, the lines that involve the DNA and form a ball'

Zeroed subjects of gerundive adverbial subclauses are also marked (7):

(7) *Essas **mudanças** podem ser para o bem ou para o mal,* `[0=<mudanças]` *atenuando sintomas de doenças ou* `[0=<mudanças]` *provocando seu desenvolvimento.* 'These changes can be for good or for evil, alleviating symptoms of disease or causing their development'

In the case of antecedent noun phrases with nominal determiners (*e.g.*, *milhão* 'million' (8) and the percentage expression *por cento* 'percent' (9) or its corresponding symbol '%'), it is the semantic head noun (syntactically, a complement of the determiner), that is chosen as the antecedent:

(8) *Segundo a última contagem do IBGE, 23,5 **milhões** de **pessoas** vivem na Amazônia.* `[0=<<pessoas]` *São apenas 13% da população brasileira, mas o suficiente para* [0=<o] *fazer um estrago de proporções planetárias.* 'According to the last count of IBGE, 23.5 million people live in the Amazon. [They] are only 13% of the Brazilian population, but enough to produce damage of planetary proportions'
(9) *Mais de 90% dos **machos** descendentes das cobaias apresentavam os mesmos problemas, sem nunca* `[0=<machos]` *terem sido expostos ao insecticida.* 'Over 90% of male descendants of the [experiment] subjects showed the same problems without ever having been exposed to insecticide'

If the head noun of a NP has been zeroed in front of determiners, the determiner is then taken as the head noun of that NP and functions as antecedent for the following zero anaphor (10); in this way, the zero anaphor always refers to its syntactic antecedent (and not to the antecedent noun itself, which can be very far way from the current sentence). This approach is also adopted for nominal determiners like *maioria* 'majority' (11):

(10) *E* ***os*** *demais, apesar de* `[0=<os]` *serem titulados, terão de ter experiência profissional na área do curso.* 'And the remaining [students], although [they] have already graduate, will have to acquire professional experience in the course's area'
(11) *Dos 25% restantes, a* ***maioria*** *pediu desculpas,* `[0=<maioria]` `explicando que [0=<maioria]` *tinha marcado de* `[0=<maioria]` *sair com a namorada.* 'From the remaining 25%, the majority appologized, explaining that [they] already had a date with their girlfriend'

The annotation of *zero indefinite subjects* is somewhat different, since they do not constitute anaphors, but may hinder significantly the anaphora resolution process. *Zero-indefinite* subjects are marked as `[0=indef]` (12):

(12) `[0=indef]` *Nascer com patrimônio genético idêntico não significa que as pessoas crescerão* `[0=<pessoas]` *tendo corpo, mente e doenças iguais.* 'To be born with identical genetic heritage does not mean that people will grow up having a similar body, mind and diseases'

First person plural indefinite subject, where there is a systematic ambiguity with zeroed pronoun *nós* 'we', is specially noted `[0=1p]`. In the example (13) , the first person plural may correspond to: (a) a real plural, referring to the speaker and his/her team of researchers; (b) the so-called 'modesty' plural, referring to the (singular) speaker; or (c) the indefinite (generic) subject, referring to the scientific community as a whole. Naturally, such ambiguities cannot be solved at this stage. Similarly, sentences with the indefinite third person plural zeroed subject, where the verb in the third person plural, is annotated `[0=3p]`, as in (14). This type of subject is systematically ambiguous between the indefinite subject and a simply zeroed third person plural pronoun *eles/elas* 'they', so that only context can disambiguate it:

(13) *As descobertas são impressionantes.* `[0=1p]` *Conseguimos informações preciosas sobre os genes, as marcas epigenéticas e as mudanças do genoma ao longo da vida, o que dá início a uma revolução.* 'The findings are impressive. We got valuable information about the genes, the epigenetic markings and the changes of the genome throughout life, which initiates a revolution'
(14) *"Ainda* `[0=3p]` *estão fazendo isso lá embaixo",* `[0=<<Zé Lopes]` *acrescenta* (...) ' "[They] are still doing it down there," [Zé Lopes] adds'

The *impersonal subject* is annotated `[0=impers]`. This notation may cover different syntactic and semantic structures, such as *meteorological* constructions (15); and *impersonal* constructions with *haver* 'to there be' (both in Brazilian and European Portuguese)(16), or *ter* 'to have' (only in Brazilian Portuguese) (17):

(15) — *Nossa!* `[0=impers] Esfriou!` '— Wow. It got cold!'
(16) *"*`[0=impers]` *Há uma perigosa tendência a* `[0=indef]` *fazer correlações entre etnia, crime e predisposição genética"* ' "There is a dangerous tendency to establish correlations between ethnic origin, crime and genetic predisposition" '
(17) `[0=impers]` *Tem gente* `[0=<gente]` *fazendo isso.* 'There is people doing this'

Finally, the subject of adjectives (and past participles when used as adjectives) is only marked if they appear with their copula verb, therefore the zeroed subjects of adjectives in apposition, as *capazes* 'capable' in (18), are not marked:

(18) *Ela ajudará na criação de* ***remédios*** *personalizados, capazes de* `[0=<remédios]` *alterar o genoma para* `[0=<remédios]` *deter o desenvolvimento de doenças e de transtornos psíquicos.* 'It will help in the creation of personalised medicine, capable of altering the genome in order to halt the development of diseases and mental disorders'

To conclude, some exceptions. *Topicalization* structures, *cleft sentences* with *ser ... que*, and other forms of focusing sentence elements involving changes in basic word-order are not marked and the syntactic position left empty by the moved constituent is not signaled. In the case of *direct speech* (for example, in interviews) the first person subject and the second person, if zeroed, are not marked. The zeroed subject of imperative sentences; direct, total (yes/no) or partial (*wh-*) interrogative sentences; question tags; and exclamative sentences are not to be marked, either. For lack of space we do not provide examples of such sentence types here.

4 Results

In this Section we present some of the main results from the annotation process. On the one hand, Table 2 presents the distribution of zero anaphors, zero-indefinite, impersonal constructions, 1p- and 3p-indefinite constructions. One can see that indefinites and impersonal constructions represent 26% of the corpus zero subjects, thus they constitute a serious hindrance to anaphora resolution. On the other hand, Table 3 shows the distribution of the anaphora/cataphora and intra-/inter-sentential distinctive types. Only 4% of the anaphoric relations correspond to instances of cataphora. The cases of intra-sentential anaphora represent 66.9% of the tags. It is noteworthy that in 53.8% cases of anaphora proper, the antecedent can not be found in the same sentence as the anaphor.

Table 2. Breakdown of zero-anaphors, impersonal and zero-indefinite subjects

ZAC corpus						
Text Type	**zero**	**indef**	**impers**	**1p**	**3p**	**Total**
Special Report	371	81	42	41	3	538
News	40	8	4	0	0	52
Chronicle	286	41	17	43	8	395
Fiction (short stories)	110	4	11	5	16	146
Fiction (novel)	281	7	26	19	25	358
Total	1,088	141	100	108	52	1,489
Total (%)		0.73	0.09	0.07	0.07	0.03

Table 3. Distribution of the anaphora/cataphora and intra-/inter-sentential anaphora.

ZAC corpus				
Text types	<	<<	>	>>
Special Reports	275	74	20	0
News	34	2	4	0
Chronicle	156	115	5	2
Fiction (short stories)	44	65	4	0
Fiction(novel)	171	99	8	0
sub-total	680	355	41	2
sub-total (%)	0.631	0.329	0.038	0.002
Total	**1,035**		**43**	
Total (%)	**0.960**		**0.040**	

5 Conclusions and Future Work

This paper presented a corpus with manually annotated zero-anaphoric relations, as well as other related phenomena with direct bearing in the anaphora resolution process of zero anaphora, namely impersonal and zero-indefinite subject constructions. A set of annotation guidelines was produced [19], and briefly presented here, to better target the linguistic phenomena and provide consistency to the annotation process. To the best of our knowledge, this is the first corpus annotated for this type of phenomena for Portuguese. Results show that zero-indefinites constitute up to $\frac{1}{4}$ of the tags, which significantly complicates the AR process, while cataphora has only less than 5% frequency. Based on this corpus, a rule-based module for anaphora resolution has already been developed by Pereira [3, 20] and integrated in the Portuguese grammar of STRING system [4]. The evaluation of this module reported a 0.60 precision, 0.46 recall and 0.52 F-measure. Later, Marques [17] developed the ARM 2.0 hybrid AR module, currently used in STRING, but only targeting pronominal anaphora.

In the future, we expect to expand the ZAC corpus in order to include European Portuguese texts and to use machine learning techniques to improve the zero anaphora resolution in STRING. We also envisage to integrate pronominal and zero anaphora phenomena into a single, unified and coherent, AR module.

Acknowledgment

This work was partially supported by the European Union Erasmus Mundus Master Courses program (ref.EMMC 2008-0083) and by national funds through FCT–Fundação para a Ciência e a Tecnologia, ref. UID/CEC/50021/2013.

References

1. Mitkov, R.: Anaphora Resolution. Pearson (2002)
2. Mendes, A.: Organização textual e articulação de orações. In Paiva Raposo, E., Bacelar do Nascimento, M., Mota, A., Segura, M., Mendes, A., eds.: Gramática do Português. Volume 2. Fundação Calouste Gulbenkian, Lisboa (2013) 1691–1755

3. Pereira, S.: Linguistics Parameters for Zero Anaphora Resolution. Master's thesis, Univ. Algarve/Univ. Wolverhampton, Faro and Wolverhampton (2010)
4. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING: A Hybrid Statistical and Rule-based Natural Language Processing Chain for Portuguese. In: 10th Conference on Computational Processing of Portuguese. PROPOR '12 (Demo Session), Coimbra, Portugal (2012) `https://string.l2f.inesc-id.pt//`.
5. Hobbs, J.R.: Resolving Pronoun References. Lingua **44** (1978) 311–338
6. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics **20**(4) (1994) 535–561
7. Dagan, I., Itai, A.: A Statistical Filter for Resolving Pronoun References. In Feldman, Y.A., Bruckstein, A., eds.: Artificial Intelligence and Computer Vision. Elsevier Science Publishers B.V. (1991) 125–135
8. Kennedy, C., Boguraev, B.: Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In: Proceedings of the 16th International Conference on Computational Linguistics. COLING '96, Copenhagen, Denmark, John Wiley and Sons, Ltd (1996) 113–118
9. Paraboni, I., Strube-de-Lima, V.L.: Possessive Pronominal Anaphor Resolution in Portuguese Written Texts. In: Proceedings of the 17th International Conference on Computational Linguistics. COLING '98, Montreal, Québec, Canada, Association for Computational Linguistics (1998) 1010–1014
10. Chaves, A.R., Rino, L.H.: The Mitkov Algorithm for Anaphora Resolution in Portuguese. In: Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language. PROPOR '08, Aveiro, Portugal, Springer-Verlag (2008) 51–60
11. McCarthy, J.F., Lehnert, W.G.: Using Decision Trees for Coreference Resolution. In: Proceedings of the 8th International Joint Conference on Artificial Intelligence. IJCAI '95, Montreal, Québec, Canada, Morgan Kaufmann Publishers Inc. (1995) 1050–1055
12. Cardie, C., Wagstaff, K.: Noun Phrase Coreference as Clustering. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. EMNLP/VLC '99, College Park, Maryland, USA (1999) 82–89
13. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics **27**(4) (2001) 521–544
14. Rahman, A., Ng, V.: Supervised Models for Coreference Resolution. In: Proceedings of Empirical Methods in Natural Language Processing. EMNLP '09, Singapore, Association for Computational Linguistics (2009) 968–977
15. Rello, L., Ilisei, I.: A comparative study of Spanish zero pronoun distribuition. In: International Symposium on Data and Sense Mining, Machine Tanslation and Controlled Languages, Besançon, France (2009) 209–214
16. Nobre, N.: Resolução de Expressões Anafóricas. Master's thesis, Universidade Técnica de Lisboa - Instituto Superior Técnico (2011)
17. Marques, J.: Anaphora Resolution in Portuguese: A Hybid Approach. Master's thesis, Universidade de Lisboa - Instituto Superior Técnico (2013)
18. Marques, J., Baptista, J., Mamede, N.: Anaphora Annotation Guidelines. Technical report, L2F - Spoken Language Laboratory, INESC-ID Lisboa, Lisboa (2013)
19. Pereira, S., Baptista, J.: Zero anaphora corpus annotation guidelines. Technical report, L2F - Spoken Language Laboratory, INESC-ID Lisboa, Lisboa (2009)
20. Pereira, S.: ZAC.PB: An annotated corpus for zero anaphora resolution in Portuguese. In: Student Research Workshop in conjunction with RANLP-09, Borovets, Bulgaria (2009) 53–59

Resources for Monolingual Translation: a case study of Text Simplification for Portuguese

Rodrigo Wilkens, Leonardo Zilio, Marco Idiart, Jorge Wagner Filho, Eduardo Ferreira, Luis Mollmann, Bianca Pasqualini, and Aline Villavicencio

Institute of Informatics, UFRGS, Brazil

{rodrigo.wilkens,lzilio}@inf.ufrgs.br,idiart@if.ufrgs.br,
{jawfilho,eduardo.ferreira,luis.mollmann}@inf.ufrgs.br,
bianca.pasqualini@gmail.com,avillavicencio@inf.ufrgs.br

Abstract. Text simplification can be seen as a monolingual translation task, and for precise results various resources are needed. Thus, in this paper we examine resources available for Portuguese and English. Among them, we discuss simple and general corpora, dictionaries of simple words, thesauri, lists of multiword expressions, and resources containing semantic role labeling. The difference in terms of quantity and coverage of manually constructed resources for these two languages reveals the gap that still needs to be addressed for Portuguese.

Keywords: lexical resources, lexical simplification, text simplification, Portuguese

1 Introduction

In text simplification (TS) the aim is to "translate" complex texts into simpler versions that maintain as much as possible the original meaning of the texts. TS can be seen as a monolingual translation task, and the quality of a TS system is often linked to the availability of resources. For instance, English abounds with resources and tools for text simplification, including the Simple English Wikipedia parallel corpus, which contains alignments between the Simple and the Standard English Wikipedia [7], and a few machine-translation-based approaches, such as [39], and also general corpora and resources like the Penn Treebank, the British National Corpus and WordNet [11]. On the other hand, for Portuguese, the limited availability of such resources presents one of the main challenges for producing good quality results.

In this paper, we aim at presenting some resources that are being used for TS, comparing them in English and Portuguese, and discussing their availability for both languages. In particular, we describe some of the resources that were created for Portuguese in order to approximate similar resources for English, and that can be used not only for TS but for other tasks and applications such as machine translation and parsing. The paper is structured as follows: we start discussing related work in TS and describing a language-independent TS

architecture (Section 2). Then we present resources developed for Portuguese, especially corpora (Section 3.1) and dictionaries (Section 3.2).

2 Related Work

Simplification has traditionally been divided into two main tasks [32]: lexical simplification (LS), which focuses on replacing complex words or expressions with simpler synonyms, and syntactic simplification (SS), which changes the structure of a sentence by using simpler syntactic constructions [33]. Recently, MT techniques have been employed for learning alignments between simple and standard sentences, so that simplification is viewed as a monolingual translation encompassing both lexical and syntactic changes [39, 7]. Additionally, explanation generation can be used for providing extra information about complex expressions without replacing them.

There are many proposals for textual simplification [1, 33, 21], which vary especially in the simplification type (e.g. lexical vs syntactic) and in the resources and tools employed. For instance, Siddharthan [33] and Aluisio et al. [1] use part-of-speech (POS) taggers and parsers along with dictionaries to perform both lexical and syntactic simplification. For LS, content word classes such as nouns, verbs, adjectives and adverbs are often targeted for substitution, and can be identified using a POS tagger and those that are considered more complex are marked for substitution. For SS, transformation rules based on specific syntactic structures are applied on the basis of parsing information.

In addition to standard dictionaries, resources for simplification include dictionaries containing paraphrases, explanations or definitions when providing complementary information about complex words and thesauri for information about synonyms when replacing complex words for simpler alternatives (e.g., *acquire* for *buy*) [17, 8, 4, 16]. Moreover, lexical substitution also involves word sense disambiguation, from standard disambiguation techniques [23] to those based on clustering [25] and a substitute ranking strategy to choose the simpler alternative, from word frequency [8, 4] to a combination of attributes [17] using machine learning [16]. In the following sections we discuss some of the resources required for these tasks focusing on Portuguese.

3 Resources for Lexical Simplification

The steps for identifying complex words and ranking alternative words are closely related to the target user and use metrics such as word and n-gram frequency, polysemy and word size [18]. These metrics are drawn from very large corpora, which could be automatically built from the Web or written with a focus on the target user. The WaC methodology resorts in a clear set of steps for creating a corpus that reflects language use. In this regard, in Section 3.1 we describe corpora that were manually built with a focus on specific target users, and some automatically constructed corpora: the brWaC [5], which has a wide coverage

of Portuguese, the WRC and readability assessed WaC, a corpus focused on readability [34].

3.1 Corpora

For English, *Simple English Wikipedia* is one of the most widely repositories of simplified language use, and Coster [7] aligned original sentences to their simplified counterparts. For Portuguese, we highlight three small corpora: (1) "Coleção É Só o Começo", described in Wilkens [37], containing five books manually simplified by linguists.[1]; (2) Caseli [6], who built a manually annotated corpus of syntactic and lexical simplifications; and (3) WikiJunior[2], illustrated books with simple, readable and friendly language for children up to 12 years old. Also, there is the corpus *Projeto PorPopular* [12], a collection of a few popular newspapers, or tabloids, read massively by low literacy readers in Brazil.

As these are valuable but small corpora, in addition to them, for building general corpora we adopt the WaCky (Web-As-Corpus Kool Yinitiative)[3] [3] framework which was successfully used to build very large general corpora for many languages, including English, German and French, with a good level of content variation and quantity of information. Using the WaCky method, a 3 billion word Portuguese corpus, the *brWaC* [5], was collected following three main steps: crawling, cleaning and near-duplicate detection and removal. For this first step, medium frequency content words are used as seeds for the crawler, according to frequency lists from the Brazilian Portuguese corpora in Linguateca[4]. The second step includes HTML and boilerplate stripping, using density metrics and shallow text features. The duplicate removal performs a pairwise comparison of all the documents retrieved, aiming to keep only distinct documents as part of the corpus. This is a crucial step, since, otherwise, corpus size may not reflect content variation. This corpus can be used as basis for building some of the resources (thesauri, language models, etc).

Focusing on low literacy speakers, we extended the pipeline of [3] with a WaC-based crawler equipped with a readability assessment module [34]. It adds an intermediate step of readability assessment between the post-crawl cleaning and the near-duplicate detection and removal. The readability assessment module is responsible for calculating several readability features for each document that are subsequently used as input to a machine learning classification model. Using this approach, two Portuguese corpora were built[5]. The first of them is the Wikilivros Readability Corpus (WRC) consisting of the HTML book library

[1] The collection is a partnership between private publishing houses with the Brazilian Ministry of Education for publishing classic literary works, most of them more than a century old, for adults with low literacy.

[2] Available at http://pt.wikibooks.org/wiki/Wikilivros:Wikijúnior

[3] `http://wacky.sslmit.unibo.it/doku.php`.

[4] Linguateca Corpora Frequency List, available at `dinis2.linguateca.pt/acesso/tokens/formas.totalbr.txt`.

[5] Both corpora are available in `http://www.inf.ufrgs.br/pln/resource/CrawlingByReadabilityLevel.zip`

from the Wikilivros Web site[6]. These books are separated in the following levels: 33 books used in the 1st to 9th grades in the Brazilian education system (from now on called *Level 1*), 65 books used in the 10th to 12th grades (*Level 2*) and 21 books used in college education (*Level 3*). The second corpus is a crawled WaC classified by readability level consisting of more than 5,000 web pages, and was used as a validation corpus. This corpus contains 129k sentences from level 1, 236k sentences from level 2 and 96k sentences from level 3, and the type-token ratio is similar in all classes (around 0.05). A clear difference in the average sentence size in words was observed among the different levels (13.5 for level 1, 15.2 for level 2 and 17.4 for level 3). The difference in proportions between the WRC and the readability-assessed WaC (the latter being almost a hundred times larger) illustrates the advantages of using automatically filtered, web-crawled content to complement manually generated materials.

3.2 Lexica and Thesauri

For lexical simplification we use 4 different types of resources: lists of simple words and thesauri for lexical substitution; semantic gold standards for evaluating distributional thesauri; Multiword Expressions lists for identifying expressions that should be treated as semantic units in text; and semantic role labeling information for helping with word sense disambiguation prior to lexical substitution.

Manually constructed resources that use lists of simple words are an attractive alternative for proposing simpler substitution candidates for simplification. In these lists, definitions are written with a controlled vocabulary using words that are considered easier to understand than the words that are defined therein [40]. For English, one such list is the Oxford 3000[7]. For Brazilian Portuguese the list of words combines 1,024 of the most common words from three corpora (Banco de Português, Dicionário Ilustrado do Português, and corpus of the tabloid Diário Gaúcho) along with the entries from Oxford 3000 translated to Portuguese, resulting in 3,853 words [13][8]. This list was manually revised and cleaned so that duplicates were removed, and expressions that would be inaccessible to the target users were also replaced with more familiar words, selected from dictionaries and the corpus *Banco de Português*.

For additional coverage, rich sources for finding synonyms for LS are manually constructed resources like WordNet [11] for English and equivalents in Portuguese like Onto.PT[9] [15], OpenWN-PT[10] [26], MultiWordnet of Portuguese[11],

[6] Wikilivros is the Portuguese version of the Wikibooks initiative (`https://pt.wikibooks.org/`)
[7] http://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000
[8] https://drive.google.com/open?id=0B0BaiG237npwaVpadHRWaDNLclU
[9] http://ontopt.dei.uc.pt
[10] https://github.com/arademaker/openWordnet-PT
[11] http://mwnpt.di.fc.ul.pt/

WordNet.PT[12] [22], WordNet.Br[13] [9], discussed in [24]. However, their availability and coverage varies, so that the resources especially for Portuguese need to be complemented by automatically constructed distributional thesauri, using language independent tools, such as Glove [27]. Given that precision-oriented lexical substitution is directly dependent on thesaurus quality, we adopt semantic gold standards for evaluating the quality of these resources. For English we use the WordNet-Based Synonymy Test (WBST) [14] which is a TOEFL like test for assessing if the thesaurus agrees with the gold standard in distinguishing between semantically associated words and distractors for a set of target words. For Portuguese we proposed a similar test, the BabelNet-Based Semantic Gold Standard (B^2SG) which evaluates nouns and verbs in three distinct relations (synonymy, antonymy and hypernymy) [38]. For each target word there is a list with 4 alternatives: one semantically related, and 3 unrelated words. For instance, for the target noun *rival* and the synonym relation, there are four alternatives: *competitor*, *curtain*, *bulwark*, and *crimson*, among which the correct alternative is the first one. The methodology we used for generating the test items ensured that the target, related and unrelated words were close in terms of frequency and polysemy. The validation of B^2SG was done first semi-automatically against a lexical resource, and for all the items not covered by the resource, two native speakers manually validated the relations. From the initial set of relations 60.7% was valid, resulting in a total of 2,875 relations in the gold standard[14].

As multiword expressions (MWEs), like idiomatic expressions and compound nouns, are a consistent source of difficulties for machine translation and lexical simplification, and are particularly affected by limited coverage in resources MWE lists are generated for English and Portuguese. For English there are initiatives like SIGLEX-MWE[15] to make available MWEs dictionaries. For Portuguese we use Europarl corpus [20] for MWE discovery adopting both a language dependent parsing-based method with the Fips "deep" linguistic parser [36, 35], and a language independent with the mwetoolkit [28]. We focused on nominal compounds of the form Noun-Preposition-Noun, and evaluate the results semi-automatically [43]. We automatically evaluated the top 2000 MWE candidates, assuming as true positives those that were found in one of the lexical resources, obtaining precision of 9.9% (Fips) and 15.5% (mwetoolkit) individually, and 17.9% combined. Additionally, given the over-strict criterion of the automatic evaluation, the top 100 candidates were manually validated, with a combination of the two methods resulting in 79.9% precision. These results indicate that a combined method produces more accurate MWE lists to be used in a text simplification tool.

[12] http://www.clul.ul.pt/clg/wordnetpt/index.html
[13] http://143.107.183.175:21380/wordnetbr
[14] The resource can be downloaded from `http://www.inf.ufrgs.br/pln/resource/B2SG.zip`
[15] http://multiword.sourceforge.net/

For lexical simplification, semantic role labeling (SRL) information can work as a means of disambiguation of words, since one word with one distinct meaning would only be selected if it could fill the criteria for the semantic role of the target complex word to be replaced. SRL resources describe the semantic function of an argument in relation to a determined word class (usually, a verb). For English resources include FrameNet [2] and PropBank [19]. For Portuguese, there are only a few initiatives for creating SRL resources: PropBank.Br [10], VerbNet.Br [30], and FrameNet Brasil [29]. VerbLexPor[16] [42] uses two different corpora: one composed of Cardiology papers (CARD), and the other by newspaper articles extracted from Diário Gaúcho (DG). Both corpora were parsed, so that syntactic information was explicit, and were then processed with a subcategorization frames extractor [44]. VerbLexPor comprises 46 semantic roles manually annotated by a linguist.[17] The resource contains semantic information on 192 verbs (77 in CARD and 191 in DG), distributed in 7,231 sentences (1,931 in CARD and 5,301 in DG), that comprise 15,281 annotated arguments (4,192 in CARD and 11,089 in DG). Besides the SRL, the resource contains syntactic information from the parser on all sentences from both corpora.

A summary of the resources for English and Portuguese is in Table1. These figures highlight the big gap in terms of resources and coverage for these two languages, especially in terms of manually constructed resources. These differences can have a direct impact in the quality of the simplification. The exception is in the automatically constructed resources which only depend on the size of the corpus.

Table 1. Resources for English and Portuguese

Resources	Size in English	Size in Portuguese
WaC	>2 billion	3 billion
Lists of simple words	3,000	1,024
WordNet-like	155,287	150,000
Semantic Gold Standard	23,570	2,875
MWE lists	71,888	3,204

4 Conclusions and Future Work

In this paper, we discussed resources for lexical simplification in English and Portuguese. These resources include manually constructed simple corpora, automatically built WaCky general corpora (ukWaC and brWaC) and corpora with readability assessment. In this regard, development of large domain- and

[16] VerbLexPor is readily available for download in XML and SQL formats: `http://cameleon.imag.fr/xwiki/bin/view/Main/Semantic\%20role\%20labels\%20corpus\%20-\%20Brazilian\%20Portuguese`

[17] The complete set of semantic roles can be found in [41].

readability-driven corpora is still an open research field. We also presented dictionaries of simple words and thesauri, but, while we have lists of simple words, there is still a need for lists of complex words, such as in [31]. In the case of multiword expression lists, we still need larger corpora from which to draw multiword information, since the corpus we used presented limitation in terms of coverage. Finally, the existing resources that display semantic role information need to be expanded and combined, so as to present more verbs and frames. Although some of these resources were produced for TS, they can also be used in tasks and applications such as parsing and machine translation.[18]

Language-independent automatic methods are an alternative for less time-consuming resource construction, as they only depend on the availability of large corpora and are developed by the whole international NLP community. However, these techniques require good methods for detecting and correcting noise in their results. For instance, distributional thesauri present related words.

As future work we there is still room for improvement in methods for synonymy detection. Besides that, we also intend to develop tools for precise morphological inflection, especially for Portuguese, since it has a richer morphology than English, but this is sometimes oversimplified.

Acknowledgments

This research was partially developed in the project *Text Simplification of Complex Expressions*, sponsored by Samsung Eletrônica da Amazônia Ltda., in the terms of the Brazilian law n. 8.248/91. This work was also partly supported by CNPq (142356/2011-5, 482520/2012-4, 312114/2015-0), CAPES (12537/12-8), and FAPERGS AiMWEst.

References

1. Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E., Fortes, R.P.: Towards brazilian portuguese automatic text simplification systems. In: Proc. of the 8th ACM symposium on Document engineering. pp. 240–248. ACM (2008)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. pp. 86–90. Association for Computational Linguistics (1998)
3. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E.: The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language resources and evaluation 43(3), 209–226 (2009)
4. Biran, O., Brody, S., Elhadad, N.: Putting it simply: a context-aware approach to lexical simplification. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 496–501. ACL (2011)
5. Boos, R., Prestes, K., Villavicencio, A., Padró, M.: brWaC: a WaCky corpus for Brazilian Portuguese. In: Computational Processing of the Portuguese Language, pp. 201–206. Springer (2014)

[18] These resources are freely available at http://www.inf.ufrgs.br/pln

6. Caseli, H.d.M., Pereira, T.d.F., Specia, L., Pardo, T.A., Gasperin, C., Aluísio, S.M.: Building a Brazilian Portuguese parallel corpus of original and simplified texts. In: Proc. of CICLing (2009)
7. Coster, W., Kauchak, D.: Simple English Wikipedia: a new text simplification task. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 665–669. Association for Computational Linguistics (2011)
8. Devlin, S., Tait, J.: The use of a psycholinguistic database in the simplification of text for aphasic readers. Linguistic databases pp. 161–173 (1998)
9. Dias-da-Silva, B.C., Felippo, A.D., das Graças Volpe Nunes, M.: The automatic mapping of Princeton WordNet lexical-conceptual relations onto the brazilian portuguese wordnet database. In: Proceedings of LREC 2008, Marrakech, Morocco. European Language Resources Association (2008)
10. Duran, M.S., Aluísio, S.M.: Propbank-Br: a Brazilian Treebank annotated with semantic role labels. In: LREC. pp. 1862–1867 (2012)
11. Fellbaum, C.: WordNet. Wiley Online Library (1998)
12. Finatto, M.J.B.: Projeto PorPopular, frequência de verbos em português e no jornal popular popular brasileiro. As Ciências do Léxico: lexicologia, lexicografia, terminologia VI (2012)
13. Finatto, M.J.B., Evers, A., Pasqualino, B.F., Kuhn, T.Z., Pereira, A.M.: Vocabulário controlado e redação de definições em dicionários de português para estrangeiros: ensaios para uma léxico-estatística textual
14. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of the Ninth Conference on Computational Natural Language Learning. pp. 25–32. Association for Computational Linguistics (2005)
15. Gonçalo Oliveira, H., Gomes, P.: Towards the automatic creation of a wordnet from a term-based lexical network. In: Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing. pp. 10–18. ACL Press (July 2010), http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira_Gomes2010_TextGraphs5_postconf.pdf
16. Horn, C., Manduca, C., Kauchak, D.: Learning a Lexical Simplifier Using Wikipedia. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (2014)
17. Kajiwara, T., Matsumoto, H., Yamamoto, K.: Selecting Proper Lexical Paraphrase for Children. In: ROCLING (2013)
18. Keskisärkkä, R., Jönsson, A.: Automatic Text Simplification via Synonym Replacement. In: Fourth Swedish Language Technology Conference (SLTC 2012), 24-26 October 2012, Lund, Sweden (2012)
19. Kingsbury, P., Palmer, M.: From TreeBank to PropBank. In: LREC. Citeseer (2002)
20. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. vol. 5, pp. 79–86. Citeseer (2005)
21. Leroy, G., Kauchak, D., Mouradi, O.: A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. International journal of medical informatics 82(8), 717–730 (2013)
22. Marrafa, P.: WordNet do Português: uma base de dados de conhecimento linguístico. Instituto de Camões, Lisboa (2002)
23. Nunes, B.P., Kawase, R., Siehndel, P., Casanova, M.A., Dietze, S.: As simple as it gets-a sentence simplifier for different learning levels and contexts. In: 2013 IEEE

13th International Conference on Advanced Learning Technologies (ICALT). pp. 128–132. IEEE (2013)
24. Oliveira, H.G., de Paiva, V., Freitas, C., Rademaker, A., Real, L., Simões, A.: As WordNets do português. Oslo Studies in Language 7(1) (2015)
25. Paetzold, G.H., Specia, L.: LEXenstein: A Framework for Lexical Simplification. ACL-IJCNLP 2015 1(1), 85 (2015)
26. de Paiva, V., Rademaker, A., de Melo, G.: OpenWordNet-PT: An open Brazilian WordNet for reasoning. In: Proceedings of the 24th International Conference on Computational Linguistics (2012), see at http://www.coling2012-iitb.org (Demonstration Paper). Published also as Techreport http://hdl.handle.net/10438/10274
27. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. vol. 14, pp. 1532–1543 (2014)
28. Ramisch, C.: Multiword Expressions Acquisition (2015)
29. Salomão, M.M.M.: FrameNet Brasil: um trabalho em progresso. Calidoscópio 7(3), 171–182 (2009)
30. Scarton, C.: VerbNet. Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. NILC/USP. Ph.D. thesis, Dissertação de mestrado orientada por Sandra Maria Aluísio (2013)
31. Shardlow, M.: The cw corpus: A new resource for evaluating the identification of complex words. ACL 2013 p. 69 (2013)
32. Shardlow, M.: A survey of automated text simplification. International Journal (2014)
33. Siddharthan, A.: An architecture for a text simplification system. In: Language Engineering Conference. pp. 64–71 (2002)
34. Wagner Filho, J., Wilkens, R., Zilio, L., Idiart, M., Villavicencio, A.: Crawling by readability level. In: Proceedings of 12th International Conference on the Computational Processing of Portuguese (PROPOR) (2016)
35. Wehrli, E., Nerima, L.: The Fips Multilingual Parser. In: Language Production, Cognition, and the Lexicon, pp. 473–490. Springer (2015)
36. Wehrli, E., Seretan, V., Nerima, L.: Sentence analysis and collocation identification. COLING (2010)
37. Wilkens, R., Dalla Vecchia, A., Boito, M.Z., Padró, M., Villavicencio, A.: Size does not matter. frequency does. a study of features for measuring lexical complexity. In: Advances in Artificial Intelligence–IBERAMIA 2014, pp. 129–140. Springer (2014)
38. Wilkens, R., Zilio, L., Ferreira, E., Villavicencio, A.: B^2SG: a TOEFL-like Task for Portuguese. Language Resources and Evaluation Conference (LREC) (2016)
39. Woodsend, K., Lapata, M.: Learning to simplify sentences with quasi-synchronous grammar and integer programming. In: Proceedings of the conference on empirical methods in natural language processing. pp. 409–420. Association for Computational Linguistics (2011)
40. Zgusta, L.: Manual of lexicography. Mouton, The Hague, Paris (1971)
41. Zilio, L.: VerbLexPor: um recurso léxico com anotação de papéis semânticos para o português. UFRGS. Ph.D. thesis, Tese de doutorado orientada por Maria José Bocorny Finatto e Aline Villavicencio. (2015)
42. Zilio, L., Finatto, M.J.B., Villavicencio, A.: VerbLexPor: um recurso léxico com anotação de papéis semânticos para o português. In: Proceedings of STIL 2015. Sociedade Brasileira de Computação (to appear)
43. Zilio, L., Wilkens, R., Santos, L., Idiart, M., Wehrli, E., Villavicencio, A.: Joining Forces for MultiwordMultiword Expression Identification. In: Proceedings of 12th International Conference on the Computational Processing of Portuguese (PROPOR) (2016)

44. Zilio, L., Zanette, A., Scarton, C.: Automatic extraction of subcategorization frames from corpora. In: New Languages Technologies and Linguistic Research: a Two-Way Road. Cambridge Scholars Publishing (2014)

Building a Brazilian Portuguese - Brazilian Sign Language Parallel Corpus using Motion Capture Data

José Mario De Martino[1], Paula D. Paro Costa[1], Ângelo Benetti[2], Luciana Aguera Rosa[3], Kate Mamhy Oliveira Kumada[3], and Ivani Rodrigues Silva[3]

[1] School of Electrical and Computer Engineering, University of Campinas, Brazil
[2] Center for Information Technology "Renato Archer", Campinas, Brazil
[3] Center of Research Studies in Rehabilitation "Gabriel Porto", University of Campinas, Brazil
martino@fee.unicamp.br

Abstract. Brazilian Sign Language, or Libras, is the language officially recognized as the first language of the Brazilian deaf community by a federal law. Nevertheless, deaf Brazilians still face considerable challenges to access public services or to advance their studies since most part of basic and advanced information is still only available in written Brazilian Portuguese (BP). In general, the knowledge of written BP by deaf citizens is far from satisfactory. In this context, automatic machine translation from BP into Libras is a promising approach to help deaf individuals to leverage their knowledge and represents a valuable option to reduce communication barriers especially in situations when a sign language interpreter is not available. This paper describes our approach to build a comprehensive BP-Libras parallel corpus. The approach combines a methodology based on the translation of school textbooks with a thorough description of sign gestures and facial expressions based on motion captured data. The methodology also seeks to handle the challenges of working with a sign language that still lacks school vocabulary.

Keywords: sign language; Brazilian sign language; machine translation; parallel corpus; signing avatar; motion capture

1 Introduction

Brazilian Sign Language (Libras) is the language used by the Brazilian deaf community. Libras as any other sign language is perceived visually and is produced by gestures composed of movements of the hand, arms and body, combined with facial expressions. Libras grammar is comprised of lexical items that structure themselves over specific morphological, syntactical and semantic mechanisms that are used as means of generating linguistic structures, allowing the production of countless phrases from a limited set of rules. In Brazil, Libras has the status of official language for the Brazilian deaf community.

According to the 2010 Brazilian demographic census, 5.1% of the population declared having at least some permanent hearing loss, representing a total of 9.7 million citizens. Among them, more than 776 thousand people were school-aged (between 0 and 17 years of age) [3]. Despite the efforts, initiatives aimed at promoting a bilingual education for these children are still sparse. Challenges range from the lack of trained professional to the absence of bilingual learning material [10]. In this context, machine translation of school textbooks written in Brazilian Portuguese (BP) into Libras is a promising approach to increase the availability of bilingual material, to promote the engagement of deaf students in classroom activities, and to improve the learning process. Considering the visual-gestural modality of Libras, we advocate the use of realistic virtual humans, or avatar, to present the results of the translation process.

In addition to educational applications, the automatic translation into Libras can also be used to assist deaf and hard-of-hearing citizens with difficulties to access written information, seeking to enhance information accessibility for deaf citizens to a level similar to that experienced by hearing ones.

Currently, we are focusing on the translation of school textbooks. A snapshot of the machine translation system that is under development is presented in Figure 1. The book, whose content can also be displayed in Libras, is shown on the left side of the figure. The student can select any written sentence of the book and see its translation into Libras displayed by the animated avatar on the right side of the figure. To support the implementation of the machine translation system, we are constructing a bilingual parallel corpus based on the translation of a series of school textbooks. This paper describes our approach to build the parallel corpus using motion capture (mocap) data.

While it is possible to identify other initiatives to construct BP-Libras corpora through the video recording of Libras interpreters [8, 9, 4], the present work contributes to the initiative of documenting BP-Libras parallel translation examples not only through annotated videos but also adding a parametric description of the signing gestures and facial expressions, through the use of mocap technology. Such approach leverages sign language data-driven research [6]. In addition, the corpus under construction is not limited to be a glossary of terms but it comprehends an extensive collection of text excerpts and complete sentences whose translations unfold Libras' grammar and structure – an important contribution for the studies on Brazilian sign language.

2 Methodology

The BP-Libras parallel corpus has been gradually collected and consolidated based on the analysis and translation of school textbooks. The corpus will be used by the machine translation system under development. As the quality of translation depends on the extent and diversity of the corpus material, our method was devised to provide a controlled and gradual expansion of the corpus through cycles of successive refinement and extension. Each cycle involves the following sequence of steps: 1) selection of a textbook (Section 2.1); 2) translation to Li-

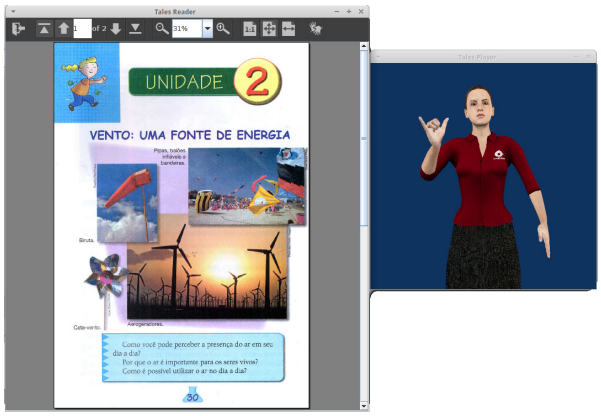


Fig. 1: Brazilian Portuguese to Brazilian Sign Language machine translation system, which presents the translation by means of a tridimensional realistic avatar.

bras of the content of the book by Libras interpreters (Section 2.2); 3) recording the Libras translation of the book in video and with motion capture equipment (Section 2.3); 4) Annotation of the videos recorded in Step 3 (Section 2.4).

2.1 Book Selection

The first step in our methodology is the definition of reference source texts to be used in the construction of the text-sign language parallel corpus. Aiming at educational applications and considering the difficulties faced by deaf children during the first years of their formal education at school, we chose to work with elementary school textbooks. In particular, we are focused on textbooks involving content related to biology, physics and geography. The selection of a textbook is based on three main criteria: the visual organization of the book; the progression of complexity in the use of the written language; and the ease of translation of the written content. Currently, we are working with an elementary science textbook for third-graders (8 or 9 year-old students). This is the first book that is being processed.

2.2 Translation by Interpreters

After the selection of the textbook, the next step of our methodology is the translation of its content by Libras interpreters. The translation process is performed collaboratively involving the close interaction between hearing native Brazilian Portuguese speakers proficient in Libras and deaf individuals having Libras as their first language and reading skills in Brazilian Portuguese. For each sentence of the textbook, a first proposal of translation is suggested by a member of the translation team. The translation proposal in the form of a provisional video is circulated among the team to get the approval of the members. As many

other living languages, Libras also presents regional, cultural and social variations. These variations are not only restricted to different signs being used to express the same concept, but also include subtle but perceptible variations in sign presentation. Seeking to guarantee an acceptable level of standardization, the proposed translations are checked against well-known Libras dictionaries as the illustrated trilingual dictionary for Brazilian Sign Language from Capovilla and Raphael [1] and the online dictionary for Brazilian Sign Language from Lira and Souza [4].

However, even with the support of standard references, it is not always possible to reach a clear consensus on the translation of the sentences. For instance, the movement for the verb MORAR (to live) found in Capovilla and Raphael's Dictionary ([1], p. 920) is accompanied by the instruction to "make the sign for *house* twice", while the Lira and Souza's dictionary indicates a single movement. Disagreement regarding hand configuration can be illustrated by the verb IMAGINAR (to imagine), executed with four fingers in Capovilla and Raphael's dictionary ([1], p. 746) and with five fingers in the Lira and Souza's dictionary. Moreover, in other cases, not only the minimal gesturing is different, but also two or three different signs are used for a single signifier. Still, the greatest challenge faced by that translation process was the lack of accepted Libras signs for specific scientific and technical terms. According to Marinho [5], one of the major hurdles for education in our context is the absence of dictionaries and glossaries in Libras listing scientific/technological terms. The most widely known Libras dictionaries in Brazil are still limited to everyday vocabulary, and the few initiatives to standardize scientific/technical glossaries have not become widespread in the deaf community yet. To overcome these problems, we performed a comprehensive research on existing field-specific Libras glossaries and sign compendia [11] and established a network of 22 individuals including deaf and hearing people, teachers and technical professionals, that regularly meet to discuss, create, and validate new scientific/technical signs. To date, more than 160 signs were created by our group. We plan on publishing and validating these signs on a nationwide basis using the Internet.

The final result of the translation process consists of a set of videos with the associated transcription using glosses [7]. Each video contains the Libras translation of a sentence of the book. Essentially, the gloss transcription transcribes in words the Libras content of the video.

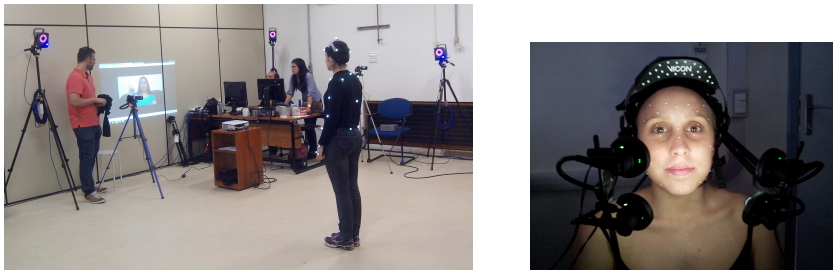
2.3 Motion Capture Sessions

The videos and glosses for each translated sentence generated in the previous step are used as reference to guide the motion capture (mocap) sessions. Mocap technology enables the capture of the movement of the body, limbs, head, and face in the tridimensional space.

We conducted the recording of the movement of a Libras interpreter in two different types of mocap sessions: one dedicated to record the gestures and the other devoted to capture facial expressions. By gestures, we mean the movement of the body, arms, and head during signing.

Figure 2(a) presents the mocap session setup to record the gestures of the translation of the sentences of the textbook. The image shows three of the eight 16-megapixel near-infrared cameras used to capture the movements of the Libras interpreter. The infrared cameras record the spatial position of spherical retro-reflective markers at 120 frames per second (fps), which are attached to the interpreter's upper body. A specific configuration of 35 markers was designed to allow the detection of pelvis, lumbar, chest, neck and head, clavicles, biceps, forearms and hands. The reconstruction of the tridimensional trajectory of the markers and the computation of the behavior of the skeleton of the interpreter was done using the software Blade of Vicon Motion Systems Ltd. Markers trajectories are stored as C3D (Coordinate 3D) standard files, and the rotation of the joints of bones of the skeleton are stored as Biovision Hierarchy (BVH) files. The animation of the avatar is generated by retargeting the captured skeleton to the avatar's bone hierarchy.

As also shown in Figure 2(a), a projection screen is positioned in front of the interpreter to help her during signing. Projection is used to show the interpreter the sentence in Portuguese, the video of the translation in Libras and the associated gloss transcription. Additionally, two traditional video cameras are also part of the session setup. One video camera is positioned in front of the interpreter and the second camera captures a side view at her right side. The videos recorded by these cameras are synchronized with the mocap data and is later annotated as described in Section 2.4.



(a) (b)

Fig. 2: (a) Motion capture setup for recording arms, body, and head movement during signing. (b) Motion capture setup to capture facial expressions.

The setup for motion capturing facial expressions is presented in Figure 2(b). For this task, we are using the Vicon Cara System, a head mounted mocap system equipped with four 1280 x 720 pixels, 60 fps, pico cameras. The system is capable of processing high resolution images and determining the tridimensional position of flat markers on the face with low signal noise and high accuracy. Although it is possible to capture signing gestures and facial expressions simultaneously, the head rig used to capture facial expressions hinders the signing performance and, in some cases, could prevent the execution of the correct signing pattern. In our

approach, the output of annotation process (Section 2.4), identifies which facial expressions are necessary to guarantee signing intelligibility. Such expressions are captured and then carefully combined with data captured from the upper body gestures during mocap post-processing phase.

2.4 Corpus Annotation

The video material recorded by front and side view cameras is synchronized with the tridimensional trajectories captured by the mocap system. Using the videos as reference, it is possible to determine the frontiers of each sign for each translated sentence. For this purpose, we are using ELAN (EUDICO Linguistic Annotator [2]) for transcription and annotation (Figure 3). Table 1 summarizes the annotation schema. The first column of the table shows the names of the tiers and the second column describes their content.

Table 1: Elan's annotation tiers.

Tier Identification	Description
Inglês (English)	English translation of the sentence.
Português (Portuguese)	Brazilian Portuguese translation of the sentence.
Libras	Sign glosses.
Sinais Compostos (Compound Signs)	Identification of compound signs. Compound signs are signs that combine a sequence of primitive signs to build a new meaning. In Libras, the sign for SCHOOL, for example, is the combination of signs HOUSE and TO STUDY.
Mão direita (Right Hand)	Right hand configuration. Number that identifies, among a finite set of possible hand configurations, the handshape, including the position of the fingers, during the period specified in the timeline.
Mão esquerda (Left Hand)	Left hand Configuration.
Expressão facial (Facial expression)	Identification of facial expressions.
FCN (Inflections, classifiers and narrative features)	Identification of verbal and nominal inflections as well as the use of narrative resources in Libras, including classifiers, which are hand movements to indicate the location and movement of objects in the description of a scene.
Comentários (Comments)	Includes comments, such as the use of dactylology (fingerspelling) or the use of signs created by our group.

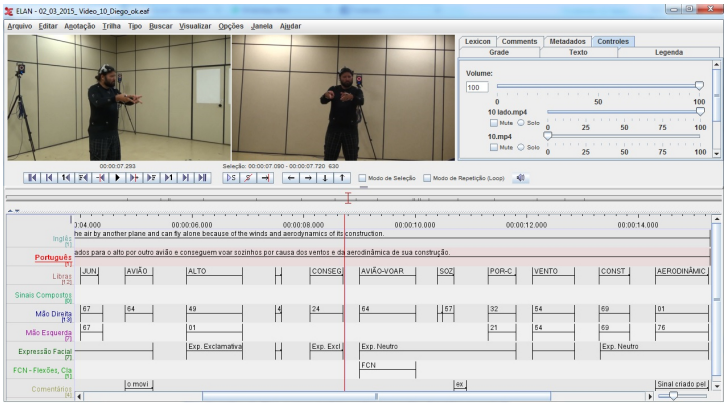


Fig. 3: Screenshot of ELAN illustrating the annotation process and highlighting the annotation tiers described in Table 1.

3 Results

The process described in Section 2 results in a Brazilian Portuguese - Brazilian Sign Language parallel corpus composed of the following set of aligned and synchronized data structures for each translated sentence:

- A frontal view video recording (MPEG-4 video file).
- A side view video recording (MPEG-4 video file).
- The signing gestures and facial expression tridimensional trajectories (C3D and BVH files).
- The corresponding ELAN Annotated Format files (EAF).

We have just finished the recording of the first school science textbook. The whole book was recorded in 23 days of mocap sessions. More than 2,000 sentences were recorded, summing approximately 8 hours of raw material.

4 Concluding Remarks

The information provided by the parallel corpus is being applied to the development of TALES, a reading assistive technology for the deaf (Figure 1).

The annotated ELAN files are being used to provide information to the animation process and to train an example-based machine translation system. It is important to note that, while the parallelism between written Brazilian Portuguese and Libras glosses can be explored by statistical machine translation approaches, the parallel text-glosses corpus does not contain enough information to guarantee sign language intelligibility after translation. In other words, the written translation to glosses lacks relevant information regarding how signs should be presented. To approach this problem, we are working on two fronts. First, we are defining an intermediate language, particularly designed to drive

the avatar animation, that will complement glosses information with relevant information for sign language synthesis, like tokens for facial expressions and other visual descriptors. Second, we are developing a mechanism to describe the signs and the transitions between them as parameters that can be analyzed by the translation algorithm towards the optimal visual sign synthesis. Additionally, it is important to emphasize the contribution of the present work to advance the studies on Brazilian Sign Language, including its systematic description and documentation.

5 Acknowledgements

This research is being supported by CNPq/MCTI-SECIS, grant number 458691/2013-5, and by Capes/SDH/MCTI, grant number 88887.091672/2014-0.

References

1. Capovilla, F.C., Raphael, W.D.: Dicionário Enciclopédico Ilustrado Trilíngue da Língua de Sinais Brasileira. Universidade de São Paulo, vol. 1 & 2. 3rd ed. (2008).
2. Crasborn, O., Sloetjes, H.: Enhanced ELAN functionality for sign language corpora. In: Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation (2008).
3. Instituto Brasileiro de Geografia e Estatística (IBGE): Censo Demográfico 2010: Características gerais da população, religião e pessoas com deficiência.
4. Lira, G.A., Souza, T.A.F.: Dicionário da Língua Brasileira de Sinais. Acessibilidade Brasil.
5. Marinho, M.L.: O ensino da biologia: o intérprete e a geração de sinais. Master's dissertation. University of Brasília (UnB), Brazil(2007)
6. Lu, P., Huenerfauth, M.: Collecting a motion-capture corpus of American Sign Language for data-driven generation research. In: Proc. of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 89-97, (2010).
7. Quadros, R.M, Karnoop, L.B.: Língua de sinais brasileira: estudos lingüísticos. Artmed Editora, (2014).
8. Quadros, R.M., Lillo-Martin, D., Chen-Pichler, D.: Methodological considerations for the development and use of sign language acquisition corpora. Spoken Corpora and Linguistic Studies 61 (2014): 84.
9. Rumjanek, V.M., Rumjanek, J.B.D., Barral, J.: Teaching Science to the Deaf: a Brazilian experience. In: INTED 2012 -6th International Technology, Education and Development Conference, 2012, Valencia. INTED 2012 -6th International Technology, Education and Development Conference, (2012).
10. Silva, I.R., Kumada, K.M.O., Hildebrand, H.R., Nogueira, A.S.: "[...] Língua de sinais eu sei, mas o português é difícil": reflexões sobre políticas lingüísticas e de identidade no contexto da surdez no fomento de práticas de letramento diferenciadas. In.: Congresso-Luso-Afro Brasileiro de Ciências Sociais, 11. Salvador. Anais. Salvador: UFBA, 2011. p. 1-16.
11. Stumpf, M.R., Oliveira, J.S., Miranda, R.D.: Glossário Letras-Libras a trajetória dos sinalários no curso: como os sinais passam a existir. In: Quadros, R.M. (eds.). Letras Libras: ontem, hoje e amanhã. Florianópolis: Edufsc (2014).