CINTIL DependencyBank PREMIUM
A Corpus of Grammatical Dependencies for Portuguese

Rita de Carvalho, Andreia Querido, Marisa Campos, Rita Valadas Pereira, João Silva, António Branco
University of Lisbon
Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
{rita.carvalho, andreia.querido, doriela.campos, ana.pereira, jsilva, antonio.branco}@di.fc.ul.pt

Abstract

This paper presents a new linguistic resource for the study and computational processing of Portuguese. CINTIL DependencyBank PREMIUM is a corpus of Portuguese news text, accurately manually annotated with a wide range of linguistic information (morpho-syntax, named-entities, syntactic function and semantic roles), making it an invaluable resource specially for the development and evaluation of data-driven natural language processing tools. The corpus is under active development, reaching 4,000 sentences in its current version. The paper also reports on the training and evaluation of a dependency parser over this corpus. CINTIL DependencyBank PREMIUM is freely-available for research purposes through META-SHARE.

Keywords: dependency bank, corpora, dependency parsing

1. Introduction

Linguistically annotated data sets are of the utmost importance for the study and computational processing of natural languages. In particular, they enable the training and evaluation of data-driven NLP tools, such as taggers, chunkers and parsers.

Over the past years, the focus of parsing tasks has moved from determining phrase constituency towards assigning syntactic function, also known as grammatical dependencies, as this sort of annotation further abstracts from the surface form of a language and is better suited for some tasks—particularly data extraction tasks—since it provides a direct representation of predicate-argument relations. As a result, data sets annotated with grammatical dependencies, or dependency banks, have also become popular (de Marneffe et al., 2006).

This paper presents CINTIL DependencyBank PREMIUM, a corpus of Portuguese newspaper text manually annotated with syntactical and semantic dependency relations, as well as part-of-speech, inflection, lemmas and named entities, making it a new important resource for NLP and for the linguistic analysis of Portuguese.

CINTIL DependencyBank PREMIUM is a new treebank, different from the existing top quality dependency bank, CINTIL-DepBank (Branco et al., 2011), in that the latter, although it uses the same tag set and annotation scheme, was built with the help of the LXGram computational grammar (Costa and Branco, 2010) and thus includes only those sentences that the grammar was able to parse, a fact that limits the coverage of that corpus to the coverage of its supporting grammar. The new CINTIL DependencyBank PREMIUM, in turn, does not have this limitation as it includes all sentences in the raw base corpus, instantiating all kinds of grammatical phenomena and lexical items that actually occur in the corpus.

CINTIL DependencyBank PREMIUM is actively developed and maintained by NLX, the Natural Language and Speech Group of the Department of Informatics of the University of Lisbon. It is available through META-SHARE.

This paper is structured as follows. Section 2. introduces CINTIL DependencyBank PREMIUM, presenting figures about the corpus and the tag set. Section 3. describes the annotation methodology and Section 4. describes the linguistic phenomena that were more interesting and challenging in terms of annotation. Section 5. reports on the experiment where a dependency parser was trained and evaluated over CINTIL DependencyBank PREMIUM. Section 6. concludes and presents some final remarks.

2. CINTIL DependencyBank PREMIUM

CINTIL DependencyBank PREMIUM is a corpus of excerpts from news articles in Portuguese, covering topics such as Economy, Politics and Sports. Therefore, it contains texts of varying lexical scope and level of grammatical sophistication. At the current stage of development, the corpus is composed by 106,590 tokens in 4,000 sentences, yielding an average of 26.65 tokens per sentence.

The annotation in CINTIL DependencyBank PREMIUM goes beyond the standard grammatical dependency relations, to include also part-of-speech tags, morpho-syntactic information (lemma and inflection) and named entity labels. Besides this, the arcs for the dependency relations between words are further decorated with tags indicating the semantic relation at stake. Table 1 shows the tag set size for each layer.

Annotation layer	Tag set size
Named entities	6
Part-of-speech	62
Inflection	105
Syntactic-semantic relations	60
grammatical dependencies	26
semantic relations	23

Table 1: Tag set size for each annotation layer

The current paper will focus on the annotation layer with the dependency relations. The syntactic dependency tag set covers the usual grammatical relation dependencies, such as Subject (SJ), Direct (DO), Indirect (IO) and Oblique (OBL) Objects, Modifiers (M), Specifiers (SP), Predicates (PRD), Complement (C), etc. The words forming a multi-word expression are connected together by a relation with the category of the expression, for instance CONJ for a multi-word conjunction.

The syntactic dependency tag set has a total of 26 tags. Some of these can be combined with semantic role tags, which total 23, yielding tags such as SJ-ARG1 (Subject, first argument) or M-CAU (Causal Modifier), resulting in the 60 syntactic-semantic dependency relation tags. For example, the syntactic tag M (Modifier) is always combined with one of 10 semantic tags, like POV (Point-of-view), TMP (Temporal) or LOC (Location). Another example are the multiple possible combinations between the SJ (Subject) tag and the semantic label indicating its role. In most cases, the Subject receives the semantic label ARG1 (first argument), but there are other labels for subjects in particular syntactic constructions, such as ARG2ac, which marks the subject in an anti-causative construction, and ARG11, which is used to mark the subject of a subject-control verb (cf. Figure 5).

3. Methodology

As a starting corpus, we take CINTIL (Barreto et al., 2006), a corpus with approximately 1 million tokens, already annotated with manually verified information on part-of-speech, morphology and named entities, and add labeled syntactic dependency relations by automatically analysing it with the LX-DepParser dependency parser.[1] This tentative annotation is then manually corrected by experts in Linguistics.

Evaluation of the automatic parsing against the outcome of the manual annotation, using the Labeled Attachment Score (LAS) metric (Nivre et al., 2007a), shows that 69.70% of the automatic annotations were assigned correctly, thus greatly reducing the amount of subsequent manual correction work that is needed.

3.1. Manual Annotation Tool

Manual correction is supported by WebAnno,[2] an open-source, general-purpose, web-based annotation system (Yimam et al., 2013). It possesses a set of design features that are useful for the annotation we need to carry out: WebAnno allows creating an annotation project and fully customize it by specifying each annotation layer in terms of its set of valid tags and type—i.e. whether the layer contains a tag per word (e.g. POS), or assigns tags to spans of words (e.g. named entities), or is composed of relations between words (e.g. dependencies). The annotated files are stored in a simple format, which allows us to convert our annotated corpus into that format and import it into WebAnno. After annotation, it is also straightforward to convert the resulting files into a standard format, such as CoNLL. The annotation process itself is supported by a user-friendly and intuitive interface that allows editing a tag by clicking on it and defining a dependency relation between words by dragging an arc between them (Figure 1 shows a sentence as viewed in WebAnno).[3] Being web-based, WebAnno runs directly in a browser, which means that it is not necessary to install any specific software on the machines used by the annotators and that all annotated files are automatically stored in the server. This is coupled with a project management design feature that allows for the administrator of a project to distribute the files to be annotated among the annotators, and a curation feature that automatically finds mismatches between annotators.

3.2. Annotation Task

To ensure a reliable linguistically interpreted data set, manual correction is done by two annotators working under a double-blind scheme, and is followed by a phase of data curation where a third annotator adjudicates any mismatches. All the annotators have graduate or postgraduate education in Linguistics or similar fields and follow specific guidelines elaborated for the task (Branco et al., 2015). Agreement between annotators, measured by taking the data produced by one of the annotators as reference and comparing it with that of the other annotator, is at 88.60% LAS.

4. Challenges

Using a statistical parser for the initial annotation enables full coverage of the corpus since every sentence gets an analysis, but it also raises some challenging issues, namely the occasional errors in the corpus that must nonetheless be annotated and the multiplicity of complex linguistic phenomena found.

4.1. Errors in the Corpus

By errors in the corpus we understand, for example, (i) punctuation errors, like placing a comma between the subject and the verb, (ii) gender and number agreement mismatches, (iii) problems involving verbal subcategorization, etc. For such mistakes we assign a special ERROR relation.

Figure 2 shows an example of a problem in the subcategorization of the verb *ajudar* (to help). In this sentence the writer incorrectly used an Indirect Object (pronoun *lhe*) instead of a Direct Object (pronoun *o*). To signal this error, the relation between the verb *ajudar* (to help) and the pronoun *lhe* (to) is labeled with the ERROR relation.

4.2. Linguistic Phenomena

Besides the ungrammatical sentences, there are multiple complex linguistic phenomena to recognize and annotate, such as comparative constructions, adverbial subordination,

[1] On-line demo at http://lxcenter.di.fc.ul.pt/services/en/LXServicesParserDep.html. For the purposes of the tentative annotation of CINTIL, the dependency parser was trained to produce syntactic-semantic labels for the relations.

[2] https://webanno.github.io/webanno/

[3] The colored boxes correspond to other annotation layers: POS (blue), lemma (yellow), inflection (red) and named entities (lilac).

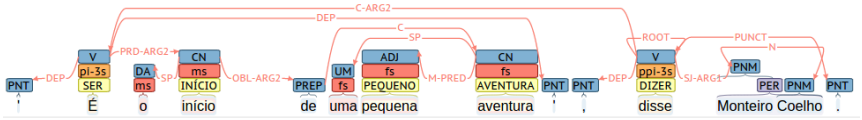

Figure 1: An annotated sentence in WebAnno: *"É o início de uma pequena aventura", disse Monteiro Coelho.* / "It is the beginning of a small adventure", said Monteiro Coelho.

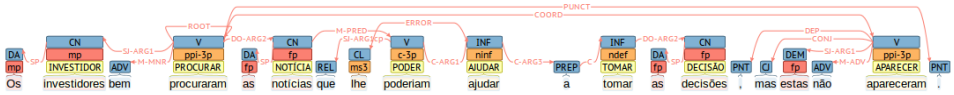

Figure 2: An ERROR relation: *Os investidores bem procuraram as notícias que lhe poderiam ajudar a tomar as decisões, mas estas não apareceram.* / The investors did look for the news that could have helped them to make the decisions, but these did not appear.

passive constructions, complex predicates (auxiliary, raising, modal and control verbs), relative constructions, null subjects, elided elements, among many others.

We have sought to ground our annotation choices on established linguistic theorizing. Sometimes, though, the grammatical phenomenon is only described in the literature without a syntactic analysis being proposed; in other cases, there are several proposed analyses that conflict with each other; and, at other times, even though we, as native speakers, may recognize a sentence as being grammatical, there are no references to the phenomenon in reference grammars such as (Cunha and Cintra, 1986), (Mateus et al., 2003) or (Raposo et al., 2013). In any case, after a thorough bibliographic research, it was necessary to make a decision. In this section we provide a few illustrative examples. For a detailed account of these decisions, please refer to the specific annotation guidelines (Branco et al., 2015) designed for the task.

4.2.1. Complex Predicates

Figure 3 shows a single sentence that contains complex predicates, involving the modal construction *deveriam ficar* (they should stay), the passive construction *foram postas* (they were placed), and the raising-to-object predicate *nos deixarem aproximar* (letting us approach).[4]

4.2.2. Emphatic Duplication

Figure 4 shows an example of emphatic duplication, i.e. when the same argument position is filled with two elements. In such cases, it is necessary to choose with which element the verb establishes its OBL-ARG2 relation. We opt for choosing the element that brings more information to the sentence. For this specific example, that is *a_ o quartel* (to the barracks) instead of *lá* (there).

4.2.3. Cleft Constructions

Another example of a phenomenon described in the literature, but for which we had to establish an analysis, is the cleft construction. One type of cleft clauses, in Portuguese, is constructed with a conjugated form of the verb *ser*+X+*que* (to be+X+that), where X is the focused expression, as exemplified in Figure 5. The constituents of the cleft construction *são... que* (it is... that) are tagged with the part-of-speech ADV and have a relation of C between them. The relation between the focused expression and the form of the verb *ser* (to be) is tagged with M-ADV.

4.2.4. Coordination with Shared Dependents

Figure 6 is an example of two coordinated elements that share the same complements or modifiers. In this example, *motivações* (motivations) and *rosto* (face) are coordinated, but the relations between these two elements and the prepositional phrase *de_ este serial killer* (of this serial killer) are different. In the noun phrase *as motivações de_ este serial killer* (motivations of this serial killer), the relation between the noun and the prepositional phrase is OBL-ARG1; while in the noun phrase *o rosto de_ este serial killer* (face of this serial killer), the relation between the noun and the prepositional phrase is M-ADV. This is represented in the dependency graph through the use of multiple heads, namely the two incoming arcs on the word *de_*, which is the head of the prepositional phrase.

5. Training a Parser

One of the main usages of linguistically annotated data sets is to support the training and evaluation of data-driven NLP tools. In this section we report on an experiment where we trained and evaluated a data-driven dependency parser over CINTIL DependencyBank PREMIUM in order to obtain an indirect indication of the quality of the dependency bank.

5.1. Representational Choices and Parsing

Dependency banks usually represent the dependency structure of a sentence through a tree whose root is the main verb of the sentence, although there are some phenomena whose linguistic analysis might be best represented through a graph structure that is not a tree, like relative structures or the coordination with shared dependents, which is represented using multiple heads.

The reason why dependency banks are usually restricted to tree structures is not linguistic, but algorithmic, since finding a parse, if such non-tree structures are allowed, can be an intractable problem (McDonald and Pereira, 2006). Though there has been some research done on extending

[4]The underscore in the token *em_* is the notation used in the CINTIL corpus to indicate that that token results from expanding a contracted form. For instance, the two tokens *em_ o* result from the tokenization stage expanding the contracted form *no*.

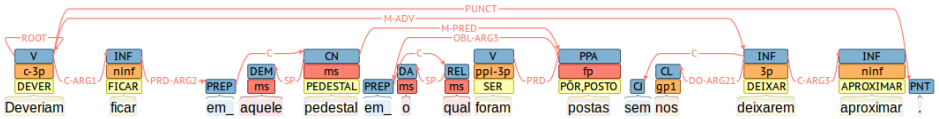

Figure 3: A complex predicate: *Deveriam ficar em_ aquele pedestal em_ o qual foram postas sem nos deixarem aproximar.* / They should stay in that pedestal in which they were placed on without letting us approach.

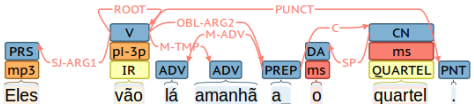

Figure 4: Emphatic duplication: *Eles vão lá amanhã a_ o quartel.* / Tomorrow they will go there to the barracks.

parsing algorithms to handle these more complex structures (McDonald and Pereira, 2006), the available dependency parsers cannot handle them. As such, these sentences are removed from the corpus before training and evaluating the parser.

In the particular case of CINTIL DependencyBank PREMIUM, 195 out of the 4,000 sentences, or nearly 5%, have to be left aside, leaving 3,805 sentences, with 98,543 tokens, to be used for this experiment.

Another representational choice that has an impact on the parsing algorithm is whether non-projective dependencies are allowed. Following the same rationale as before, given that some linguistic phenomena are best represented by non-projective dependencies, the guidelines do not prevent their use. This is not a serious impediment to the choice of parser since there are many parsing algorithms (e.g. (McDonald et al., 2005)) that are able to handle graphs with non-projective dependencies.

Out of the 3,805 sentences used in the experiment, approximately 9.5% have non-projective dependencies.

5.2. Universal Dependencies

Universal Dependencies (UD) is an annotation scheme that strives to be usable and consistent across multiple languages. It is based on Stanford Dependencies (de Marneffe et al., 2014) and it further incorporates the universal POS tag set of Petrov et al. (2012) and the Interset (Zeman, 2008) morpho-syntactic tag set.

CINTIL DependencyBank PREMIUM does not follow the UD annotation scheme. Recognizing that UD is adopted by many in the NLP community as a *de facto* standard, we have developed a tool for converting the CINTIL annotation scheme to UD, similar to what is done for multiple languages in the HamletDT collection of corpora (Zeman et al., 2014).

5.3. Tag Set Granularity

CINTIL DependencyBank PREMIUM has a highly detailed tag set of dependency labels due to syntactic relations having been extended with their semantic roles, as described in Section 2..

We run two experiments over the corpus, one with the full syntactic-semantic tag set, and another where the semantic role is removed from the dependency labels, yielding *strictu sensu* dependency relations.

5.4. Parser Evaluation

We used MaltParser (Nivre et al., 2007b), a generic parsing framework whose components can be configured by the user. These components are the parsing algorithm, which works over state transitions; the module that builds the feature vector representation of the parser state; and a Support-Vector Machine classifier that, given a parser state, chooses a parser action.

The MaltParser system allows tuning any of these three components. For the purposes of this experiment, we rely on MaltOptimizer (Ballesteros and Nivre, 2012), a tool that automatically runs and tests MaltParser under different preset configurations, and picks the setup that provides the best LAS score.

MaltOptimizer, after running a 5-fold cross-validation experiment over the data set, finds a MaltParser configuration that gives 80.77% LAS when using the full tag set, and 83.12% LAS with the coarser tag set without semantic roles, as summarized in Table 2.

dependency tag set	LAS
with semantic roles	80.77%
without semantic roles	83.12%

Table 2: Parser evaluation

As expected, performance is lower when the more granular tag set is used, due to the inevitable data-sparseness issues that come from using a larger tag set.

The LAS score when assigning only syntactic function labels (i.e. without a semantic role) is below the best results that have been achieved for Portuguese in similar tasks, like the 87.6% achieved by the best system on the CoNLL-X Shared Task[5] (Buchholz and Marsi, 2006), or the 88.24% reported in (Silva, 2014). We note, however, that these other systems were obtained by training over larger corpora, viz. 9,100 sentences for the former and 5,400 sentences for the latter.

[5]While the parser trained on CINTIL DependencyBank PREMIUM is worse than the best system in the CoNLL-X shared task, it scored well above the average of the participating systems, which is at 80.6% LAS.

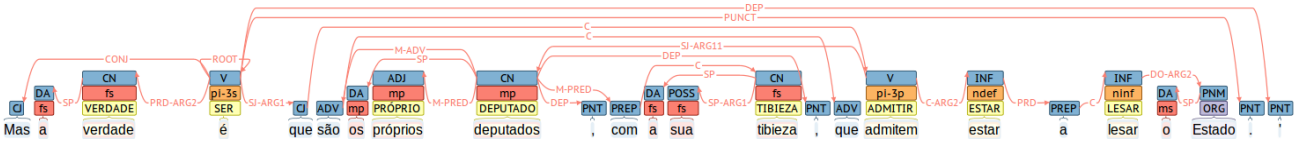

Figure 5: Cleft construction: *Mas a verdade é que são os próprios deputados, com a sua tibieza, que admitem estar a lesar o Estado.* / But the truth is that it is the deputies themselves that, with their own indifference, admit to be damaging the State.

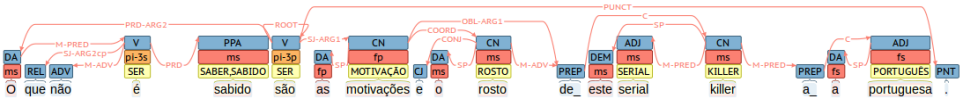

Figure 6: Coordination with shared dependents: *O que não é sabido são as motivações e o rosto de_ este serial killer a_ a portuguesa.* / What is not known are the motivations and the face of this Portuguese-style serial killer.

6. Final Remarks

CINTIL DependencyBank PREMIUM is under active development, our goal being to keep expanding it until all sentences from the underlying corpus have been annotated. As a growing corpus with a large variety of annotated complex linguistic phenomena, CINTIL DependencyBank PREMIUM can be used for training statistical dependency parsers that are to be used in a broad range of applications that need to deal with unrestricted text. At its current size, with 4,000 sentences, it already allows training a dependency parser that achieves the very competitive score of 83.12% LAS, a value that is bound to increase as the corpus grows.

CINTIL DependencyBank PREMIUM brings a new annotation layer into the existing corpus. Given the breadth of linguistic phenomena that are represented, it enables further linguistic studies that need to search that corpus for specific dependency structures.

CINTIL DependencyBank PREMIUM is available through the META-SHARE repository.[6]

7. Acknowledgments

This work was partly funded by the Portuguese Foundation for Science and Technology through the Portuguese project DP4LT (PTDC/EEI-SII/1940/2012) and by the European Commission through project QTLeap (EC/FP7/610516).

8. Bibliographical References

Ballesteros, M. and Nivre, J. (2012). MaltOptimizer: A system for MaltParser optimization. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, pages 2757–2763.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M. F., Nunes, F., and Silva, J. (2006). Open resources and tools for the shallow processing of Portuguese: The TagShare project. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*, pages 1438–1443.

Branco, A., Castro, S., Silva, J., and Costa, F. (2011). CINTIL DepBank handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2011-03, University of Lisbon.

Branco, A., Silva, J., Querido, A., and de Carvalho, R. (2015). CINTIL DependencyBank PREMIUM handbook: Design options for the representation of grammatical dependencies. Technical Report DI-FCUL-TR-2015-05, University of Lisbon.

Buchholz, S. and Marsi, E. (2006). CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Natural Language Learning (CoNLL)*, pages 149–164.

Costa, F. and Branco, A. (2010). A deep linguistic processing grammar for Portuguese. In *Proceedings of the 9th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*, number 6001 in Lecture Notes on Artificial Intelligence (LNAI), pages 86–89. Springer.

Cunha, C. and Cintra, L. (1986). *Nova Gramática do Português Contemporâneo*. Edições João Sá da Costa, 3rd edition.

de Marneffe, M., MacCartney, B., and Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*, pages 449–454.

de Marneffe, M., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., and Manning, C. (2014). Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 4585–4592.

Mateus, M. H. M., Brito, A. M., Duarte, I., Faria, I. H., Frota, S., Matos, G., Oliveira, F., Vigário, M., and Villalva, A. (2003). *Gramática da Língua Portuguesa*. Caminho, 5th edition.

McDonald, R. and Pereira, F. (2006). Online learning of approximate dependency parsing algorithms. In *Proceedings of the 11th European Chapter of the Association for Computational Linguistics (EACL)*, pages 81–88.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the 10th Con-*

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

ference on Empirical Methods in Natural Language Processing (EMNLP), pages 523–530.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., and Yuret, D. (2007a). The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the 11th Conference on Natural Language Learning (CoNLL)*, pages 915–932.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007b). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, pages 2089–2096.

Eduardo Paiva Raposo, et al., editors. (2013). *Gramática do Português*. Fundação Calouste Gulbenkian: Lisboa.

Silva, J. (2014). *Robust Handling of Out-of-Vocabulary Words in Deep Language Processing*. Ph.D. thesis, University of Lisbon.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the Demo Session at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2014). HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC)*, pages 213–218.