Texto Livre: Linguagem e Tecnologia
E-ISSN: 1983-3652
revista@textolivre.org
Universidade Federal de Minas Gerais
Brasil

http://periodicos.letras.ufmg.br/index.php/textolivre
Belo Horizonte, v. 11, n. 3, p. 1-25, set.-dez. 2018 – ISSN 1983-3652
DOI: 10.17851/1983-3652.11.3.1-25

MORPHOBR: AN OPEN SOURCE LARGE-COVERAGE FULL-FORM LEXICON FOR MORPHOLOGICAL ANALYSIS OF PORTUGUESE

MORPHOBR: UM LÉXICO DE FORMAS PLENAS DE AMPLA COBERTURA PARA A ANÁLISE MORFOLÓGICA DO PORTUGUÊS

Leonel Figueiredo de Alencar
Universidade Federal do Ceará
leonel.de.alencar@ufc.br

Bruno Cuconato
EMAp/Fundação Getúlio Vargas
bcclaro@gmail.com

Alexandre Rademaker
EMAp/Fundação Getúlio Vargas and IBM Research
arademaker@gmail.com

ABSTRACT: One of the prerequisites for many natural language processing technologies is the availability of large lexical resources. This paper reports on MorphoBr, an ongoing project aiming at building a comprehensive full-form lexicon for morphological analysis of Portuguese. A first version of the resource is already freely available online under an open source, free software license. MorphoBr combines analogous free resources, correcting several thousand errors and gaps, and systematically adding new entries. In comparison to the integrated resources, lexical entries in MorphoBr follow a more user-friendly format, which can be straightforwardly compiled into finite-state transducers for morphological analysis, e.g. in the context of syntactic parsing with a grammar in the LFG formalism using the XLE system. MorphoBr results from a combination of computational techniques. Errors and the more obvious gaps in the integrated resources were automatically corrected with scripts. However, MorphoBr's main contribution is the expansion in the inventory of nouns and adjectives. This was carried out by systematically modeling diminutive formation in the paradigm of finite-state morphology. This allowed MorphoBr to significantly outperform analogous resources in the coverage of diminutives. The first evaluation results show MorphoBr to be a promising initiative which will directly contribute to the development of more robust natural language processing tools and applications which depend on wide-coverage morphological analysis.

KEYWORDS: computational linguistics; natural language processing; morphological analysis; full-form lexicon; diminutive formation.

RESUMO: Um dos pré-requisitos para muitas tecnologias de processamento de linguagem natural é a disponibilidade de vastos recursos lexicais. Este artigo trata do MorphoBr, um projeto em desenvolvimento voltado para a construção de um léxico de formas plenas abrangente para a análise morfológica do português. Uma primeira versão do recurso já está disponível gratuitamente *on-line* sob uma licença de *software* livre e de código aberto. MorphoBr combina recursos livres análogos, corrigindo vários milhares de erros e lacunas. Em comparação com os recursos integrados, as entradas lexicais do MorphoBr seguem um formato mais amigável, o qual pode ser compilado diretamente em transdutores de estados finitos para análise morfológica, por exemplo, no contexto do *parsing* sintático com uma gramática no formalismo da LFG usando o sistema XLE. MorphoBr resulta de uma combinação de técnicas computacionais. Erros e lacunas mais óbvias nos recursos integrados foram automaticamente corrigidos com *scripts*. No entanto, a principal contribuição de MorphoBr é a expansão no inventário de substantivos e adjetivos. Isso foi alcançado pela modelação sistemática da formação de diminutivos no paradigma da morfologia de estados finitos. Isso possibilitou a MorphoBr superar de forma significativa recursos análogos na cobertura de diminutivos. Os primeiros resultados de avaliação mostram que o MorphoBr constitui uma iniciativa promissora que contribuirá de forma direta para conferir robustez a ferramentas e aplicações de processamento de linguagem natural que dependem de análise morfológica de ampla cobertura.

PALAVRAS-CHAVE: linguística computacional; processamento de linguagem natural; análise morfológica; léxico de formas plenas; formação de diminutivos.

1 Introduction

Morphological analysis is a prerequisite for syntactic parsing with linguistically motivated formalisms like LFG (BUTT et al., 1999; FALK, 2001; DIPPER, 2003), HPSG (POLLARD; SAG, 1994), and GF (RANTA, 2011). Moreover, it is useful in a wide range of applications ranging from dictionary lookup in e-book readers and spell-checkers to sentiment analysis, opinion mining, information extraction, and text classification algorithms. It maps word forms to all possible representations consisting of lemma and grammatical features, which may be filtered by subsequent processing steps (BEESLEY; KARTTUNEN, 2003; DIPPER, 2003; JURAFSKY; MARTIN, 2009).

As Alencar et al. (2014, p. 59-60) observes, "finite-state transducers, due to compact storage and fast processing, have been a preferred implementation of morphological analyzers". A finite-state transducer (henceforth FST) is a two-tape automaton. In the case of a morphological analyzer, the first tape encodes the analysis (or parse) strings and the second tape the surface strings, i.e. word forms (BEESLEY; KARTTUNEN, 2003).

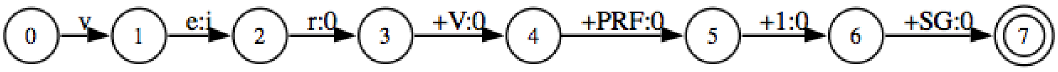

Figure 1: Example of a single-path FST.

Figure 1 exemplifies a simple FST relating the analysis string *ver+V+PRF+1+SG* to the surface string *vi* 'I saw', first person singular perfect indicative tense of Portuguese verb *ver* 'see'. This FST has 8 states connected by 7 arcs, constituting a single path from the initial state 0 to the final state 7. Each path in an FST represents an analysis string with its corresponding surface string. The arcs are labelled by symbol pairs of the form *x:y*, where *x* is a symbol of the analysis string and y is a symbol of the surface string. Labels of the form *x:x* are simplified to *x*. Label 0 represents the empty string.

Large lexical transducers with millions of paths were compiled for many languages, using a variety of finite-state tools (BEESLEY; KARTTUNEN, 2003; HULDEN, 2009), and used in diverse applications, for example in industrial-scale LFG grammars for deep syntactic parsing with the Xerox Linguistic Environment (XLE) (BUTT et al., 1999; DIPPER, 2003). One additional advantage of FSTs is that they are inherently bidirectional devices, so the same FST can serve both as a generator and an analyzer.

The construction of a wide-coverage morphological analyzer is a continuous task that goes through successive refinements. Ideally one starts from available electronic dictionaries consisting of several hundreds of thousands of word-parse pairs, so-called full-form lexicons, that can be compiled directly into an FST. However, one difficulty in reusing existent resources is that they usually adopt incompatible annotation schemes or fail to provide information that is needed for a full-fledged morphological analyzer. Besides, these resources may have errors and inconsistencies. Although many such deficiencies can be automatically corrected using simple text processing techniques, overcoming some of these problems may be less than trivial.

A more challenging task is expanding the coverage of existent dictionaries in order to deal with new words. Finite-state morphology is the standard approach to tackle this problem (BEESLEY; KARTTUNEN, 2003). It facilitates the computational modeling of grammatical regularities underlying inflection and word formation processes. From these models FSTs can be compiled for analysis and generation of new lemmas and word forms (ALENCAR et al., 2014).

In this paper we present MorphoBr, a full-form lexicon constructed from the combination, revision, and expansion of available free analog resources for Portuguese, mostly derived from Label-Lex (ELEUTÉRIO et al., 1995) and Unitex-PB (MUNIZ, 2004). This effort is part of a research project that aims at developing wide-coverage computational grammars for deep parsing of Portuguese texts. A first version of the resource is already freely available online under an open source, free software license[1].

The main advancement of MorphoBr in relation to previous resources is a finite-state component that almost triples the inventories of nouns and adjectives by computationally modeling the formation of diminutives with *-(z)inh-.* This is one of the most productive derivational processes in Portuguese. Diminutives mainly function as a means of expressing speakers' emotions and attitudes. As such, they are very common in emotive discourse. Therefore, processing diminutives should be a basic capability of systems dealing with sentiment analysis, opinion mining, etc.

In the next section, we briefly introduce the previous resources that were combined into MorphoBr. In section 3 we first give an overview of the present stage of MorphoBr, describing the format used and explaining the conversion from the source formats. We also detail errors and inconsistencies that we found in the resources and how we solved them. In section 4 we report on the finite-state implementation of diminutive formation in Portuguese. Section 5 presents evaluation experiments, showing that MorphoBr outperforms the resources it integrates. The last section sums up the main results and points out directions for further research.

[1] URL: https://github.com/LFG-PTBR/MorphoBr

2 Full-form lexicons for Portuguese

LABEL-LEX is a collection of resources.[2] According to Eleutério et al. (1995), the LABEL-LEX dictionaries originate from Costa and Melo (1991). In Ranchhod, Mota, and Baptista (1999), the authors describe a new version of the system, composed of a lexicon and grammars. The lexical data are organized according to the formal complexity of the lexical units. In the current version, the lexicon is organized in three different files: LABEL-LEX-sw (inflected forms), LABEL-LEX-mw (multi-word forms), and LABEL-LEX-gr (grammars)[3]. Only the first one is freely available for download. The LABEL-LEX-sw version 4.1 contains 938,445 word-parse pairs, e.g. *gato,gatar.V:P1s*. In this entry, the lemma *gatar* 'to fail' and the verb category label *V* are assigned to the word form *gato*, classifying it as first person singular of the indicative present.

LABEL-LEX resources were expanded (with adaptations) and integrated in many different systems over time. To mention some of them UNITEX[4], INTEX[5], and FreeLing (PADRÓ; STANILOVSKY, 2012). In Garcia and Gamallo (2010), the authors describe the integration of Label-Lex dictionary into FreeLing and its conversion to the EAGLES tag schema (LEACH; WILSON, 1996). Besides the tags conversion, many other adaptations were done to mitigate conflicts between morphological decisions of the dictionaries and of the data collected from annotated corpora. The current Portuguese dictionary in FreeLing version 4.0 is composed of 1,214,093 word forms.

Later, Garcia et al. (2014) presents different dictionaries of the new orthography as well as a new freely available testing corpus, containing different varieties and textual typologies. The combined European and Brazilian Portuguese dictionary expanded the FreeLing dictionary with more 492,896 word-parse pairs. Considering all the additions and improvements to the Label-Lex distribution made by Garcia and Gamalo, we opted to use the Garcia and FreeLing files (henceforth GFL) instead of the Label-Lex distribution.

Following the European projects, the Brazilian NILC group developed their own lexical resource in the context of the Unitex-PB project (MUNIZ, 2004)[6]. It is divided into four files: DELAS-PB (single-word lemmas), DELAF-PB (inflected forms), inflectional graphs, and DELACF-PB (multi-word forms). Version 2 from May 2015 of DELAF-PB contains 9,072,338 word-parse pairs, of which more than 8 million are verbs, 80,000 are nouns, and 90,000 are adjectives.

Both FreeLing and DELAF-PB are freely distributed under free software, open source licenses. FreeLing dictionaries are obtained from different open-source external projects, so they have different copyright holders and different licenses than the rest of FreeLing packages. Nevertheless, both FreeLing's Portuguese dictionary and DELAF-PB are freely distributed under free software, open source licenses. The former is available under the General Public License 3, the same license of the original LABEL-LEX distribution[7]. The latter is distributed under the GNU Lesser General Public License 2.1[8],

[2] URL: http://label.ist.utl.pt/pt/labellex_pt.php

[3] Only the first one is still freely available for download.

[4] URL: http://unitexgramlab.org

[5] URL: http://intex.univ-fcomte.fr

[6] URL: http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/

[7] URL: http://nlp.lsi.upc.edu/freeling/index.php/node/1

[8] URL: http://www.gnu.org/licenses/lgpl-2.1.html

which is approved by both the Free Software Foundation[9] and the Open Source Initiative[10]. This means that end-users can freely use these libraries and developers can modify them and share the modified versions.

These two resources differ in how they handle derived words. While Unitex-PB lists all diminutives with -*(z)inh*- in DELAF-PB, FreeLing 4.0 applies affixation rules to analyze formations like *dorzinha* at runtime (i.e. during text analysis), correctly classifying them as diminutive forms of the corresponding lemma (*dor* 'pain' in the case at hand). However, FreeLing's dictionary contains more than 5,000 -*(z)inh*- diminutives derived from nouns and adjectives, including hundreds of completely regular formations (e.g. *abelhinha*, *abertinho*, and *amorzinho*, derived from *abelha* 'bee', *aberto* 'open', and *amor* 'love', respectively) and cases with predictable stem alterations (e.g. *amiguinho* from *amigo* 'friend' and *aneizinhos* from *anel* 'ring' in plural), which could be analyzed by means of such rules. This shows that FreeLing does not consistently handle regular diminutive formation in Portuguese by means of rules as a means to spare storage space.

3 The resource

The present version of MorphoBr comprises the most numerous word classes, namely, nouns, adjectives, adverbs, and verbs. The other classes will be incrementally added in the next development stages. The major part of the data derives from the corresponding entries from DELAF-PB, which were not only converted to the format described below, but also corrected, enriched, or supplemented. The rest of the data consists of the following sets: (i) the set difference between GFL and DELAF-PB and (ii) the expansions described in section 4.

Tokenization decisions affect what counts as candidate for entry in a full-form lexicon. This is specially the case with verb-clitic clusters in Portuguese, which in principle can be treated as a single unit or tokenized into two different units, each one with its own entry in the dictionary. In Portuguese, clitic pronouns can occur in three contexts: (i) left-adjacent to the verb in *proclitic* position, (ii) right-adjacent to the verb in *enclitic* position, or (iii) between infinitive base form[11] and inflectional endings in *mesoclitic* position, as exemplified in (1), (2), and (3), respectively. In these interlinear glosses, as in analogous examples in this paper, we use the Leipzig Glossing Rules[12], the most widespread conventions for glossing linguistic examples. While the clitic is separated from the verb by a blank in proclitic position, in enclitic and mesoclitic position it is separated by one and two hyphens, respectively.

(1) nos visitávamos
nos=visitávamos

[9] URL: http://www.gnu.org/licenses/license-list.html#SoftwareLicenses

[10] URL: https://opensource.org/licenses

[11] The *-ar* ending of the verb form in mesoclitic clusters corresponds historically to the infinitive ending (VILLALVA; SILVESTRE, 2016, p. 138). Elsewhere, however, the *a* segment functions as thematic vowel and *r* is part of the tense-mood morpheme, compare (3) with *nos visitaremos* 'we will visit ourselves'.

[12] URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

REFL;1PL;ACC=visit:IMPF;1PL
'we visited ourselves'
(2) visitávamo-nos
visitávamo=nos
visit:IMPF;1PL=REFL;1PL;ACC
'we visited ourselves'
(3) visitar-nos-emos
visitar=nos=emos
visit=REFL;1PL;ACC=FUT;1PL
'we will visit ourselves'

In cases where the pronominal clitic is not separated from the verb by a blank, we adopted DELAF-PB's strategy of handling the verb-clitic cluster as a single token with its own entry in the dictionary, as exemplified in (4) and (5), glossed in (2) and (3). GFL adopts a different strategy in this regard, assigning clitic and verb form to different entries. Thus, when converting data from GFL to MorphoBr, verb entries like (6) were discarded, because this verb form is only used with an enclitic pronoun, as in (2). The canonical form is the one in example (1).

(4) visitávamo-nos,visitar.V+PRO:I1p
(5) visitar-nos-emos,visitar.V+PRO:F1p
(6) visitávamo visitar VMII1P0

The dictionary entry format used in MorphoBr corresponds to the standard format in finite-state morphology (BEESLEY; KARTTUNEN, 2003). It consists of pairs of the form *(w, p)* separated by a NEWLINE, where *w* is a word form and *p* is an analysis string, separated by a TAB. The analysis string consists of a series of tags, the first one being the lemma. The other tags represent morphosyntactic features[13]. Tags are separated by a plus sign (a dot is used to separate the features of clitics, see below). In (7), the lemma *comprar* 'to buy' is assigned to the verb form *comprei*, classified as first person singular of the indicative perfect tense:

(7) comprei   comprar+V+PRF+1+SG

The tagset used in MorphoBr was designed to be more mnemonic than the formats of the two resources it draws upon. For this reason, tags are mnemonic abbreviations, generally following conventions common in the linguistic literature. Almost all tags were drawn from Fradin (s.d.), whose tagset represents an improvement in relation to the one of the Leipzig Glossing Rules[14]. By contrast, many tags in DELAF-PB, as shown in Table 1, are completely arbitrary single-letter abbreviations which are difficult to decode.

Table 1: Comparison between tags from DELAF-PB and MorphoBr.

[13] The complete documentation of the tagset is available in the project repository.

[14] SBJR (for present subjunctive) and PQP (for pluperfect) are among the few exceptions. In this case, the corresponding categories are missing in Fradin (s.d.).

DELAF-PB	**MorphoBr**	**Meaning**
W	INF	infinitive
K	PASSPT	passive participle
J	PRF	perfect indicative
S	SBJR	present subjunctive
T	SBJP	past (imperfect) subjunctive
U	SBJF	future subjunctive
Y	IMP	imperative

Tags in FreeLing are less arbitrary compared to DELAF-PB. However, FreeLing departs from widely accepted assumptions in linguistics. First, it represents mood and tense by two different tags. However, in the morphological analysis of Portuguese, these two types of information are usually represented by one single inflectional suffix (MONTEIRO, 1987; ROCHA, 2008; VILLALVA; SILVESTRE, 2014).[15] Second, the left-to-right order of the tags in FreeLing does not always reflect the concatenation of the morphemes representing the individual morphosyntactic features. This is the case with the diminutive forms of nouns, as in example (8) from FreeLing. In these forms, gender and number tags F and P (feminine and plural, respectively) precede diminutive tag D. However, in the surface form *abelhinhas* (diminutive of *abelha* 'bee'), glossed in (9), the diminutive suffix *-inh-* precedes the inflectional endings representing feminine plural.

(8) abelhinhas abelha NCFP00D
(9) abelh-inh-a-s
bee-DIM-F-PL

In entries for diminutive forms of adjectives, however, the tags in FreeLing do reflect morpheme concatenation order, as can be seen in (10). In this example, tag C, which represents the diminutive morpheme of adjectives, precedes the gender and number tags.

(10) abertinhas aberto AQCFP00
(11) abert-inh-a-s
open-DIM-F-PL

Differently than FreeLing, MorphoBr handles the diminutive of nouns and adjectives uniformly, since there is no morphological difference between the formation of diminutives in these two word classes. In MorphoBr, the order of tags encoding morphosyntactic features of nouns, adjectives, and verbs directly reflect the order of the corresponding

[15] MorphoBr represents person and number of verb forms by different tags, although these constitute one single morpheme according to Monteiro (1987), among others. The reason for this discrepancy resides in the mapping of morphological tags onto subject agreement features in the syntax. Both finite verb forms and adjective forms encode number information, but only the former encode person information. This shows person and number to be independent from one another.

morphemes, as exemplified in (12) and (13)[16]:

(12) abelhinhas abelha+N+DIM+F+PL
(13) abertinhas aberto+A+DIM+F+PL
(14) comprá-va-mos comprar+V+IMPF+1+PL
buy-IMPF-1PL

One of the main goals of MorphoBr is morphological analysis with FSTs in the context of syntactic parsing. The entry format adopted allows for straightforward conversion to the so called spaced-text format exemplified in (15), where word characters and tags are separated by a blank. This format, in turn, can be compiled into an FST with the proprietary Xerox Finite-State Tools (XFST) (BEESLEY; KARTTUNEN, 2003) or with Foma (HULDEN, 2009), its free-software, open source clone, using the command *read spaced-text*. In (15), the first line maps to the upper (analysis string or lexical) language of the transducer, while the second line maps to the lower (or surface string) language (BEESLEY; KARTTUNEN, 2003).

(15)
c o m p r a r +V +PRF +1 +SG
c o m p r e i

The format used in MorphoBr also allows for the enrichment of the annotation of verb-clitic clusters such as (2) and (3), since DELAF-PB provides no lemma or morphosyntactic information for clitics in enclitic or mesoclitic position, as evidenced by (4) and (5), where the *+PRO* tag only specifies that the verb is conjoined with a clitic. Clitic properties, however, are crucial for deep syntactic parsing and semantic analysis in formalisms like LFG, because the clitic realizes an argument of the verb (FALK, 2001). Therefore, in MorphoBr, the *+PRO* tag from DELAF-PB is substituted by one of our conversion tools for a sequence of tags representing the grammatical properties of the clitic.

Table 2: Comparison between DELAF-PB'S and MorphoBr's annotation of verb-clitic clusters (tags in italics and explanations in quotes).

DELAF-PB	*.V*	*+PRO*	“clitic pronoun”	*I3p*	“imperfect indicative 3rd person plural"
MorphoBr	*+V*	*.ele* ‘he’	“lemma”	*+IMPF+3+PL*	“imperfect indicative 3rd person plural”
		.ACC *.3* *.M.PL*	“accusative” “3rd person” “masculine plural”		

[16] In Portuguese, as in other inflectional languages (as opposed to agglutinative languages such as Turkish), it is often the case that different morphosyntactic features are collapsed in one single affix. For example, in *compro* 'I buy', suffix -*o* represents both present indicative and first person singular.

	+V	*.nós* 'we'	"lemma"	*+IMPF+3+PL*	"imperfect indicative 3rd person plural"
		.AD *1.PL*	"accusative or dative" "1st person plural"		

This conversion is not always a simple string replacement operation, because clitic *nos* can be ambiguous in relation to lemma depending on the verb form. While this clitic unambiguously represents accusative or dative of *nós* 'we' in entries like (4) and (5), it can be lemmatized in a two-fold way in entries like like (16), namely, either as *nós* 'we' or as *eles* 'they'. Our conversion tool successfully handles all these verb-clitic clusters, assigning the clitic in each case the corresponding lemma and inflectional informations. Accordingly, entry (16) is converted to the two entries (17) and (18), where the lemma ambiguity is resolved: in (17), the lemma of the clitic is *ele* 'he', while it is *nós* 'we' in (18).[17]

(16) compravam-nos, comprar.V+PRO:I3p
(17) compravam-nos   comprar+V.ele.ACC.3.M.PL+IMPF+3+PL
(18) compravam-nos   comprar+V.nós.AD.1.PL+IMPF+3+PL

Table 2 shows the correlation between the different components of the annotation of entries (16)-(18). Tags describing properties of clitics are separated by a dot instead of the plus sign used for the other tags. This distinction is relevant for syntactic processing. For example, *.PL* and *+PL* both encode plural number, but the former specifies the number of an object of the verb, while the latter specifies the number of the subject.[18]

Two aspects of the annotation scheme adopted for verb-clitic clusters in cases like (17) and (18) should be highlighted. First, the relative order of the individual tags describing properties of clitics is somewhat arbitrary, because these elements consist of portmanteau morphemes, conflating different properties in one single morphological unit. For example, clitic *nos* conflates case, person, and number in one single morpheme. A clitic such as *las*, however, can be segmented into different gender and number morphemes, motivating the tag sequence *F.PL.*, as exemplified in (19).

The second aspect refers to the the position of clitic tags as a whole in relation to inflectional tags of the verb. Due to direct translation of tag sequence *.V+PRO* from DELAF-PB in the way specified in Table 2, clitic tags precede verb inflectional tags. In this case, annotation does not mirror morpheme concatenation, since the enclitic pronoun follows the verb. In mesoclitic clusters, however, the clitic precedes the inflectional suffixes of the verb, see (20), so that morpheme concatenation is obeyed. We leave for a future version of MorphoBr a solution to this discrepancy.

[17] Entry (18) is ambiguous between accusative and dative case. In Portuguese, this case ambiguity pervades the pronominal system: the formal distinction between accusative and dative is neutralized in all clitic pronouns except those of the 3rd person, e.g. accusative form *o* and dative form *lhe* of *ele* 'he'. To take this state of affairs into account, we collapsed dative and accusative case of 1st and 2nd person clitics into the composite tag *.AD*, thereby preventing the duplication of hundred thousands of entries. The annotation of clitics is further explained below.

[18] See Dipper (2003) for a detailed explanation on how morphological tags (e.g. *+DIM*, *+M*, *.M*, *+PL* , *.PL* etc.) are converted to syntactical constraints in an LFG grammar implemented in the XLE system.

(19) compramo-las   comprar+V.ele.ACC.3.F.PL+PRF+1+PL
(20) comprar-lhe-emos   comprar+V.ele.DAT.3.SG+FUT+1+PL

For the conversion between the different formats, two tools were developed and are made available in the project repository. The first tool is an ad hoc Python script that converts from DELAF-PB's format to MorphoBr's. The second tool adopts a more systematic approach. It performs conversion between DELAF-PB's, GFL's format and MorphoBr's using a declarative approach. However, it can not enrich the annotation of clitics yet.

Using the GF programming language, we wrote a set of grammars that specify how abstract trees (encoding the dictionaries entries data) are mapped to strings in each dictionary's concrete format. GF grammars specify both parsers and printers, so we are able to use them to translate between different formats by parsing a string in a certain format to an abstract tree, and then printing it in another format (with some preprocessing to handle the spacing).

In the example below, we can see a GFL-formatted entry being parsed to an abstract syntax tree in the first output line. The following lines show the linearization of this abstract syntax tree in MorphoBr's format, which in this case corresponds to two entries, one for each gender (forms that admit two genders are collapsed into one "common gender" in the GFL representation).

(21)

```
Morpho> import MorphoMBR.gf MorphoFL.gf
Morpho> parse -tr -lang=FL "dentistas dentista N C C P 0 0 0"
        | linearize -all -lang=MBR
mkN "dentistas" "dentista" (mkNF Common Pl ZDegree)
dentistas       dentista +N +M +PL
dentistas       dentista +N +F +PL
```

Table 3: Examples of incongruences in the encoding of homonymous verb forms in DELAF-PB (incomplete paradigms lack some or all forms).

Types of homonymous forms	**Lexical representations**	**Total number of entries**	**Paradigm coverage**
comprava	1st person singular of imperfect indicative of the 1st conjugation	12714	complete
	3rd person singular of imperfect indicative of the 1st conjugation	**0**	**incomplete**
vendia	1st person singular of imperfect indicative of the 2nd conjugation	800	complete
	3rd person singular of imperfect indicative of the 2nd conjugation	800	complete
comprara	1st person singular of pluperfect of the	12714	complete

	1st conjugation		
	3rd person singular of pluperfect of the 1st conjugation	**5**	**incomplete**
vendera	1st person singular of pluperfect of the 2nd conjugation	800	complete
	3rd person singular of pluperfect of the 2nd conjugation	800	complete
comprar	1st person singular of future subjunctive of the 1st conjugation	12714	complete
	3rd person singular of future subjunctive of the 1st conjugation	12714	complete
vender	1st person singular of future subjunctive of the 2nd conjugation	800	complete
	3rd person singular of future subjunctive of the 2nd conjugation	**72**	**incomplete**

While converting DELAF-PB's entries to our format, we found numerous errors and inconsistencies in the original data. Just a few problems of this sort were detected when converting GFL's data. All detected problems were corrected either by the format conversion tools or by shell scripts. Due to space limitations, we limit ourselves here to the more significant examples.

DELAF-PB has 318,683 repeated entries. It contains 176 simple formatting errors due to spurious colons in inappropriate places, compare the incorrectly formatted entries in (22a) and (23a) with the corrected counterparts in (22b) and (23b), respectively.

(22) a. abstinhas:-lhe,abster.V+PRO:I2s
b. abstinhas-lhe,abster.V+PRO:I2s
(23) a. mantinhas:,manter.V:I2s
b. mantinhas,manter.V:I2s

There are 3070 verb-clitic clusters without the obligatory separator hyphen, e.g. *pruirlhes,pruir.V+PRO:W3s* instead of *pruir-lhes,pruir.V+PRO:W3s*. In 714 entries, the last character is number *1* instead of letter *s* (singular), e.g. *abstrair,abstrair.V:W31*. Another problem are 26,151 systematically missing verb forms: 1st conjugation verbs lack 3rd person singular forms of imperfect indicative (all forms missing) and pluperfect (all but 5 missing), while most 2nd conjugation verbs lack the 3rd person singular of future subjunctive (728 forms missing from a total of 800). In all these cases, the missing 3rd person singular forms are identical to the corresponding 1st person singular forms, as exemplified in Table 3.

One could argue that these forms were omitted following a general design decision to spare storage space. However, the evidence counters this assumption, suggesting that the omissions are unexpected side-effects of the lexicon compilation process. In fact, DELAF-PB's documentation does not refer to this lexicon-size reduction strategy. If it were implemented, there should be a special tag for collapsing the person information of the homonymous forms. Instead, the usual tag combination *1s* for 1st person singular is used. Moreover, as Table 3 shows, DELAF-PB does explicitly encode both members of similar pairs of homonymous forms, such as 1st and 3rd person singular forms of both imperfect indicative and pluperfect of 2nd conjugation verbs.

4 Expanding lexical coverage by means of FSTs

Diminutive formation is one of the most productive derivational processes in Portuguese (ROCHA, 2008, p. 122-123). In this language, there are more than 20 diminutive suffixes (CUNHA; CINTRA, 1985). These suffixes underlie lexicalized words from different lexical categories, ranging from nouns and adjectives to adverbs, numerals, personal pronouns, and verbs (RIO-TORTO, 2016, p. 359). They are continually used for the creation of new words. In fact, every noun and every adjective can have a diminutive form (LAPA, 1982, p. 77-82; ROCHA, 2008, p. 123; LIMA, 2011, p. 137). Among the diminutive suffixes, -*inh*- and -*zinh*- are the most productive (LAPA, 1982, p. 79; CUNHA; CINTRA, 1985, p. 91-92; RIO-TORTO, 2016, p. 373). This section describes a finite-state implementation of the formation of diminutives from these two suffixes in Portuguese. The goal of this implementation was expanding the coverage of MorphoBr, since the integrated resources have tens of thousands of gaps in this regard.

Denotatively, diminutive suffixes form hyponyms from nouns or modulate the intensity of adjectives, but more often they just convey speaker's emotions or attitudes, e.g. appreciation, dislike, empathy, politeness, etc. (LAPA, 1982; CUNHA; CINTRA, 1985; ROCHA, 2008; BAZENGA, 2012; VILLALVA; SILVESTRE, 2014; RIO-TORTO, 2016), being very common in emotive discourse, opinion or persuasive texts, etc. Therefore, analyzing diminutives should be a basic capability of NLP systems targeted at sentiment analysis, opinion mining, text classification, etc.

Out of the total of 361,485 nouns and adjectives converted from DELAF-PB and GFL (henceforth DGFL-ADJN), only 15,938 are diminutives, 10,338 of which are formed with -*(z)inh*-. Consequently, there are many gaps in these resources, because, for tens of thousands of words, they do not provide the corresponding diminutive. In DELAF-PB, for example, there are diminutives for *cobra* 'snake', *jacaré* 'alligator', *zebra* 'zebra', *gavião* 'hawk', and *cheiro* 'smell', but not for *elefante* 'elephant', *javali* 'boar', *jumento* 'donkey', *amor* 'love', *odor* 'smell', and *dor* 'pain'. Another deficiency of DELAF-PB and GFL is the lack of -*zinh*- for corresponding -*inh*- diminutives. For example, both resources include *cobrinha* (diminutive of *cobra* 'snake'), but not the equally grammatical parallel form *cobrazinha.* All these gaps seem completely arbitrary. In fact, the corresponding diminutives can easily be found in texts on the Internet, e.g.:

(24) cobrazinha, amorzinho, dorzinha, elefantinho, elefantezinho, jumentinho, jumentozinho

To fill these gaps, we took the standard assumption that the lexicon of a natural language consists not only of existent words, but also includes potential words, i.e. words that can be created by applying word-formation rules to existent words (ROCHA, 2008;

VILLALVA; SILVESTRE, 2014). Finite-state morphology is the standard paradigm for the construction of rule-based computational models of inflectional and word-formation processes. This approach has two strengths. First, morphological processes can be formalized in a way that closely mirrors linguistic descriptions. Thus, one does not have to reinvent the wheel when implementing a certain morphological phenomenon already described in detail in the linguistic literature. All one needs to do is to translate the description from a natural language into a formal specification. Second, this formal specification can be compiled into an FST using free, open source software, e.g. Foma (HULDEN, 2009). The resulting FST, in turn, can be used in a compact and efficient way during text processing.

Our implementation of diminutive formation with -*inh*- and -*zinh*- generally follows the analysis by Villalva and Silvestre (2014) and Rio-Torto (2016), according to which there are two types of diminutive suffixes in Portuguese: evaluative suffixes (-*inh*-, -*it*-, etc.) and z-evaluative suffixes (-*zinh*-, -*zit*-, etc).

As Villalva and Silvestre (2014, p. 119-120) points out, there is some variation in the distribution of these two types of suffixes across different dialects of Portuguese. A typical case of regional divergence are -*inh*- diminutives like *anelinho* (from *anel* 'ring'), derived from stems ending in *l* in singular (with a null thematic marker surfacing as *e* in plural), which are restricted to European Portuguese. Besides, there are preference differences between speakers or depending on word length or frequency. Diminutives from longer, less frequent words like *parlamento* 'parliament' are preferably constructed with z-evaluative suffixes. Accordingly, while both *parlamentozinho* and *parlamentinho* are grammatical, the former is considered more acceptable.

Abstracting away from these factors, the generalization holds that evaluative suffixes are restricted to stems of words ending with one of the thematic unstressed vowels -*o*, -*a*, and -*e* (e.g. *cheiro*, *zebra*, and *elefante*), while z-evaluative suffixes, as exemplified in Table 4, attach to inflected words. Therefore, for the first group of words, both types of suffixes are licensed (cf. *cobrinha* and *cobrazinha* from *cobra* [ˈkɔbrɐ] 'snake'), while all other words only license z-evaluative suffixes (compare *cafezinho* and **cafeinho* from *café* [kɐˈfɛ] 'foot' or *motorzinho* and **motorinho* from *motor* 'motor').

Table 4: Examples of diminutives with -*zinh*- derived from plural forms.

Singular base form	**Plural base form**	**Singular and plural diminutives**
menino 'boy'	*meninos*	*meninozinho meninozinhos*
menina 'girl'	*meninas*	*meninazinha meninazinhas*
flor 'flower'	*flores*	*florzinha florezinhas*
luz 'light'	*luzes*	*luzinha luzezinhas*
alemão 'German'	*alemães*	*alemãozinho alemãezinhos*
azul 'blue'	*azuis*	*azulzinho azuizinhos*

Apparent exceptions to this generalization are due to adjustments in the orthographic shape of the concatenated elements, due to general orthographic or phonological constraints in the language, e.g. *lapisinho* from *lápis* 'pencil' can be explained by deletion of *z* from *-zinh-*. Other exceptions are lexicalized words like *colherinha* (diminutive of *colher* 'spoon'). In deviation from the productive pattern in the standard language, the *-inh-* suffix in this example applies to a base ending in a phoneme other than one of the thematic vowels *-o*, *-a*, and *-e*. Examples like **reporterinho* (from *repórter* 'reporter') evidence that this pattern is not productive.

In the two-level approach in finite-state morphology, morphological regularities are factorized into two modules (BEESLEY; KARTTUNEN, 2003). The morphotactics component describes the possible combinations of morphemes. In the second module, morphographemic alternations handle allomorphy, i.e. changes in the orthographical form of morphemes when combined to build new words or word forms (ŠEVČÍKOVÁ, 2018)[19]. While some alternations reflect phonological changes affecting pronunciation, as in *lapisinho* referred to above, other alternations are purely orthographical, as exemplified in Table 5.

Table 5: Examples of orthographical changes not affecting pronunciation.

lemma	**base stem**	**diminutive**	**stem alomorph**	**orthographical change**
beiço 'lip'	*beiç* [bejs]	*beicinho*	*beic* [bejs]	*ç=>c*
amigo 'friend'	*amig* [ɐ'mig]	*amiguinho*	*amigu* [ɐ'mig]	*g => gu*
faca 'knife'	*fac* [fak]	*faquinha*	*faqu* [fak]	*c => qu*

Following the two-level approach, we modeled the morphotactical phenomena in the formation of *-(z)inh-* diminutives by means of a grammar in the LEXC formalism. This a high-level declarative language that enables the specification of finite-state automata and transducers in a linguistically intuitive way (BEESLEY; KARTTUNEN, 2003). Both XFST and Foma, its free software, open source counterpart, provide compilers for LEXC.

A LEXC grammar specifies different classes of elements, modeled as LEXICONS, and the possible combinations between them, as in the oversimplified example in (25). This example generates the *-(z)inh-* diminutive forms of *alegre* 'happy'. The first line declares the symbols that constitute multicharacter arc labels in the FST the grammar compiles into. The following lines specify five LEXICONS, each consisting of one or more entries. The first element in each entry is a pair of the form *x:y*, where *x* is a fragment of the analysis string and *y* is a fragment of the surface string. Typically, this pair represents a morpheme, i.e. an atomic unit of meaning and form. The identity relation *x:x* can be encoded as *x*. The empty string is represented by 0. The second element in an entry is a *continuation class*, which defines the LEXICON the *x:y* pair can be concatenated with. The number sign "#" represents the end of a word. This grammar compiles into an FST relating analysis strings like *alegre+A+DIM+F+PL* to their corresponding surface forms, *alegre^inh^a^s* and *alegre^zinh^a^s* in this case, where the caret "^" represents a morpheme boundary. These are intermediate forms which are converted to the actual forms *alegrinhas* and *alegrezinhas* by conditional replacement rules in the module encoding morphographemic alternations, see Beesley and Karttunen (2003) for a detailed explanation.

[19] In this paper, we restrict ourselves to the orthographical representation of words and morphemes. The two-level approach, however, is not restricted to this level. It can be also applied to phonological representations (BEESLEY; KARTTUNEN, 2003).

(25)

```
Multichar_Symbols +A +DIM +M +F +SG +PL
LEXICON Root
alegre Lemma ;
LEXICON Lemma
+A:0   Suffix ;
LEXICON Suffix
+DIM:^inh      Gender ;
+DIM:^zinh         Gender ;
LEXICON Gender
+M:^o Number ;
+F:^a  Number ;
LEXICON Number
+SG:0# ;
+PL:^s          # ;
```

Long-distance dependencies between elements of non-adjacent classes can be elegantly modeled by means of unification-based constraints named *flag diacritics* (BEESLEY; KARTTUNEN, 2003). These constraints function as filters at runtime during parsing and generation, blocking ungrammatical paths. Obtaining the same effects in a LEXC grammar without resorting to flag diacritics makes the code less intuitive. Alternatively, as Beesley and Karttunen (2003, p. 339) points out, excluding ungrammatical paths by composition of transducers can make the size of the resulting transducer explode.

The first class in our grammar contains 334,284 noun and adjective entries from the total of 361,485 in DGFL-ADJN. Seemingly improbable bases were filtered out, e. g. diminutives (cf. **lapisinhozinho*) and superlatives (cf. **rapidissimozinho* from *rapidíssimo* 'very fast'). Augmentatives, however, were included, since diminutives derived from these formations are considered grammatical (RIO-TORTO, 2016, p. 364) and attested in texts on the Internet (e.g. *casarãozinho* from *casarão*, augmentative of *casa* 'house'). The other classes of the LEXC grammar contain the diminutive suffixes, the gender morphemes, and the number morphemes.

Diminutive formation in Portuguese involves a dependency relation between the gender of the base and the gender marker, as exemplified in (26)-(29). This dependency is non-local because the diminutive suffix intervenes between these two elements.

(26) a. problem-inh-a
problem(M)-DIM-M.SG
b. problem-a-zinh-o
problem-M.SG-DIM-M.SG
(27) a. trib-inh-o
tribe(F)-DIM-F.SG
b. trib-o-zinh-a
tribe-F.SG-DIM-F.SG
(28) a. pont-inh-a

bridge(F)-DIM-F.SG
b. pont-e-zinh-a
bridge-F.SG-DIM-F.SG
(29) a. dent-inh-o
tooth(M)-DIM-M.SG
b. dent-e-zinh-o
tooth-M.SG-DIM-M.SG

Examples (26)-(29) show that gender marking in Portuguese diminutives constitute a complex phenomenon, because masculine gender can also be marked by *a* and feminine gender by *o*, which are the canonical markers for feminine gender and masculine gender, respectively, compare (26a) and (27a) with (26b), (27b), (28), and (29). Another difficulty for the formalization is the opposing behaviour of -*inh*- and -*zinh*- in cases such as (26) and (27): while the former selects the non-canonical markers, the latter selects the canonical markers. In our LEXC grammar, these facts are handled by means of flag diacritics in a linguistically intuitive way.

Morphographemic alternations were modeled by a cascade of 11 conditional replacement rules. These rules follow the general template *A --> B C _D*, which informally reads "substitute A for B in the context of C and D", where C is the left-hand and D the right-hand context. This module includes, besides the orthographical changes from Table 5, plural -*s* deletion before *z* (*azuis+zinhos => azuizinhos*, plural diminutive of *azul* 'blue'), thematic vowel deletion (*casa+inha => casinha,* diminutive of *casa* 'house'), optional thematic *e* deletion in plurals like *luzezinhas* and *florzinhas* (producing *luzinhas* and *florzinhas*) and stem *i* deletion before another *i* (e.g. *cheiinho=>cheinho* and *saiinha=>sainha*, derived from *cheio* 'full' and *saia* 'skirt', respectively).

The current implementation is biased towards contemporary Brazilian Portuguese, where the forms generated by the last two rules are attested in standard language texts and considered grammatical[20]. Another consequence of the present limitation is that the formation of diminutives like *anelinho*, which is only productive in European Portuguese, was not implemented yet.

Both the LEXC grammar and the morphographemic alternations were compiled into FSTs, which, in turn, were composed into one single FST, which we call DIM1[21]. This FST has almost 2 millions paths, but just a fraction is licensed by the unification constraints that operate during analysis or generation. In fact, extracting the grammatical word-parse pairs from DIM1 reveals that their number amounts to 625,716 pairs (see Table 6). In order to reduce complexity, as proposed by (ALENCAR et al., 2014), these word-parse pairs where compiled into a second FST using the read spaced-text command referred to in section 3. This derived FST we designate by DIM2. Among other reasons, DIM2 is less complex than DIM1 because there are no unification constraints to be solved.

[20] On these forms, see Cipro Neto (s.n.t.) and Nogueira (2010). Double i deletion is not restricted to diminutive formation, as evidenced by superlative forms like *seríssimo*, derived from *seriíssimo* 'very serious' (CUNHA; CINTRA, 1985, p. 251).

[21] Commented source code as well as all test sets referred to below are available in MorphoBr's repository. The code compiles with both Foma and XFST.

Table 6: Comparison of the complexity of different FSTs compiled with Foma[22]. The second column specifies space in disk of the word-parse pairs and the third, FST size in memory.

FST	**Description**	**File**	**Memory**	**States**	**Arcs**	**Paths**
DIM1	composition of LEXC grammar and morphographemic rules	22M	2.8M	78336	185891	1958232
DIM2	compilation of all word-parse pairs from DIM1	22M	2.8M	77903	184804	625716
ADJN	all word-parse pairs from DGFL-ADJN	10M	2.8M	80428	184975	361485
ALL	union of DIM2 and ADJN	32M (220)	3.6M (28.6)	93334 (16)	237254 (28.3)	977071 (170.3)

Table 6 compares the complexity of DIM2 to two other FSTs, which we label ADJN and ALL. ADJN was compiled by applying the read spaced-text command to all word-parse pairs from DGFL-ADJN (i.e. the set of all nouns and adjectives that were converted from DELAF-PB and GFL). ALL is the FST resulting from union of DIM2 and ADJN. The last line of Table 6 allows us to assess coverage gain as well as the cost of uniting DIM2 with ADJN. The numbers in brackets show the percentage increases in relation to the numbers in the penultimate line. ALL has 170,3% more paths than ADJN, but the complexity cost in terms of FST size and number of arcs and states ranges from 16% to 28,6%. On the other hand, the word-parse pairs from ALL occupy 220% more space in disk than those from ADJN.

5 Evaluation

In order to assess coverage and accuracy of our resource, experimental evaluation was carried out in two different phases of the project development. In the first phase, evaluation was restricted to MorphoBr's entries from DELAF-PB and GFL. Three experiments were performed in this stage. First, these entries were used to lemmatize the Universal Dependencies (henceforth UD) Portuguese GSD corpus. In this way, we could have a measure of the resource's coverage on real-world data. In the next experiment, an accuracy test was carried on Bosque (another Portuguese UD corpus), by comparing the lemmas MorphoBr assigned to the words in this corpus to the ones Bosque actually had. In the third experiment, we measured FreeLing's coverage of verb clitic-clusters by means of suffix rules.

[22] All transducers were also successfully compiled with XFST.

In the second project development phase, evaluation was restricted to the diminutive entries created by the finite-state approach and encoded in DIM2. Coverage comparisons were performed against the two previous relevant resources: (i) the set of -*(z)inh*- diminutives from DGFL-ADJN, i.e. the set of all nouns and adjectives converted from DELAF-PB and GFL, and (ii) FreeLing's suffix rules.

5.1 Improving GSD

The UD Portuguese GSD corpus was converted from the Google UD Treebanks (MCDONALD et al., 2013) and it is now part of the UD project (NIVRE et al., 2017). Even though the GSD corpus is officially an UD corpus, it is still considered incomplete, partly because it does not contain the lemmas for its words[23]. Following the UD guidelines, we corrected the annotation of *$* and *%*, which are wrongly tagged as NOUN instead of SYM in the original corpus. Given a word form and the PoS tag of a token, we converted the word form to lower-case and searched for the pair in MorphoBr. In this experiment, we only consider the grammatical classes which are already part of our resource, i.e., verbs, adjectives, adverbs, and nouns. The tokens with the remaining UD PoS tags were ignored.

The results of the experiments are classified in three cases: (i) *missing*: 2.8% of the tokens have no lemma in MorphoBr (e.g. *km*, *cerca*, *quarta*, *torcida*, *mensalão*); (ii) *unique*: 93% of the tokens have a unique lemma in MorphoBr; (iii) *multiple*: 4.2% of the tokens have more than one possible lemma (e.g. *foi*, *foram*, *era*). These results can be partly explained by the many differences between the corpus annotations, UD guidelines, and MorphoBr's design decisions. For instance, *quarta* 'fourth' and *terceiro* 'third' are considered determinants in MorphoBr, not numerals as in the UD guidelines[24]. On the other hand, many missing lemmas are abbreviations such as *TV* and *Km* or parts of multi-word expressions (MWEs) such as *a cerca de* 'regarding to', which are presently not handled by MorphoBr. We also identified a few misspellings in the corpus. The true missing words from MorphoBr are cases such as the nouns *torcida* 'supporters' (or 'cheering') and *mensalão* 'big monthly stipend' (neologism derived from the augmentative of adjective *mensal* 'monthly')[25].

5.2 Bosque comparison

We also compared our dictionary coverage to the UD-Portuguese-Bosque corpus. Its original lemmatization was provided by the PALAVRAS system (BICK, 2014) and it was manually revised. As in section 5.1, we used the pair (word form,PoS) to look up the appropriate entries in MorphoBr. For each such pair, there could be a single value, multiple values, or no values at all. We classified each of the 87,623 tokens in the corpus in four cases: (i) 80,614 tokens (92%) have the *same lemma*, only one possible value for the (word form,PoS) pair and it matches the lemma on the corpus; (ii) 1,025 tokens (1.16%) have a *different lemma*, that is, the lemma in the corpus differ from the lemma in MorphoBr; (iii) for 3,664 (4.1%) tokens MorphoBr contains *more than one value* (and thus

[23] URL: https://github.com/universaldependencies/UD_Portuguese-GSD

[24] URL: http://universaldependencies.org/u/pos/NUM.html

[25] URL: https://www.economist.com/the-economist-explains/2013/11/18/what-is-brazils-mensalao

cannot be automatically compared to the lemma in the corpus); (iv) 2,320 tokens (2.6%) where *missing* in MorphoBr.

Generally the issues with the difference between lemmas is due to MWEs and proper noun tokenization, and also divergences between how to lemmatize words, as each dictionary adopts a particular stance. For example, in the original Bosque corpus MWEs were tokenized as a single unit (e.g. *em termos* 'in terms'). UD guidelines suggest that these expressions should be split into tokens connected via the *fixed* dependency. It seems that, when Bosque was converted to UD, the split was indeed performed, but the MWE was not consistently lemmatized (e.g. 'em termos' was lemmatized as *em* and *termos*, instead of *em* and *termo*). There are also divergences in how MorphoBr lemmatizes certain words and how lemmatizations were done in Bosque. E.g., the adjective *maior* 'greater' is lemmatized in MorphoBr as *maior* but in Bosque as *grande* 'big'. Nouns such as *filha* 'daughter' are lemmatized in MorphoBr as *filho* 'son', but in Bosque as *filha*.

5.3 Comparison to DGFL-ADJN

In this section we evaluate our finite-state implementation of diminutive formation with -*(z)inh-,* mainly comparing it to DGFL-ADJN*.* Comparison to FreeLing's affix rules is the subject of section 5.4*.* Here, we first perform quantitative comparisons. Next, we evaluate the implementation qualitatively, in order to assess to what extent it is linguistically correct, in that the diminutives generated are grammatically well-formed. One way to do that would be to ask human experts to provide grammaticality judgements. However, since DIM2 contains more than half a million diminutives, a manual evaluation seems not to be practical.

We have seen in section 4 that DGFL-ADJN only contains 10,338 word-parse pairs with -*(z)inh-* diminutives, while DIM2 encodes 625,716 such pairs, which represents an increase of 5,952.58%. Since diminutives in DIM2 were generated from noun and adjective bases from DGFL-ADJN, it contains pairs that are already part of DGFL-ADJ. However, DIM2 has 615,586 new -*(z)inh-* diminutives pairs. This amounts to an increase of 5,854.59% or 59,5 times. On the other hand, there are only 208 such pairs in DGFL-ADJN which are not contained in DIM2. These include lexicalized irregular or dialectally restricted formations as well as errors in DGFL-ADJ, as we will see below.

A qualitative evaluation of a finite-state implementation of a morphology fragment involves two aspects: (i) whether the FST generates the correct forms from given analysis strings and (ii) whether it provides the correct analysis strings for the given surface forms. As evidenced by cases like *cheinho*, *lapisinho*, and *alemãezinhos*, commented on in section 4, distribution of -*(z)inh-* as well as allomorphy depend on properties of the bases, e.g. thematic class, stem-final or word-final grapheme, etc. Therefore, to test generation from DIM2, we first manually compiled TEST-UP, comprising 208 analysis strings like *luz+N+DIM+F+PL*, representing the diminutives of the different types of noun and adjective bases in Portuguese. These types mainly derive from the exhaustive classification of thematic classes of nouns and adjectives from Villalva and Silvestre (2014), but cases discussed in Cipro Neto (s.n.t.), Nogueira (2010), and Rio-Torto (2016) are also included. For TEST-UP, DIM2 generated the expected word forms.

Next, we tested both generation and analysis in comparison to DGFL-ADJN. To this end, we extracted analysis strings and word forms of all -*(z)inh*- diminutives from DGFL-ADJN, resulting in the test sets D-UP with 9,303 items and D-LOW with 9,372 items, exemplified in (30) and (31), respectively. DIM2 was then applied to both test sets, attaining 99,7% and 98,4% coverage, respectively. The strings that were not recognized by DIM2 are *missing items*. For example, DIM2 does not produce any analysis for the word form *coleginha* nor does it generate any surface form for the analysis string *cebola+N+DIM+M+SG*, so these strings are labeled missing items.

(30) abalado+A+DIM+F+PL
abalado+A+DIM+F+SG
abalado+A+DIM+M+PL
abalado+A+DIM+M+SG
(31) abaladazinha
abaladazinhas
abaladinha
abaladinhas
abaladinho
abaladinhos

Let us now see why these two types of items were not recognized, which caused DIM2 to fall short of 100% coverage in the two test sets. Of the 150 missing items from D-LOW, 88 are due to deviations from the standard language in DGFL-ADJN: 67 contain orthographic errors, see (32), and 21 violate standard rules of diminutive formation, see (33).

(32) **lebõezinhos* (*lebrõezinhos*), **coleginha* (*coleguinha*), **portuguezinho* (*portuguesinho*), **avózinha* (*avozinha*), **carcaçinha* (*carcacinha*), **paíszinho* (*paisinho*)
(33) **azulzinhas* (*azuizinhas*), **alemãozinhos* (*alemãezinhos*), **probleminho* (*probleminha*)

A total of 12 missing items contain double *ii* in diminutives, e.g. *cheiinho* (from *cheio* 'full'). While these forms are standard in European Portuguese (RIO-TORTO, 2016, p. 364), in present Brazilian Portuguese, the corresponding variants with a single *i* are preferred.

Most other missing items are irregular diminutives or result from dialectally restricted formation processes. Many of these forms are lexicalized. There are 32 -*inh*-diminutives from stems ending in *r*,*l*, or *u*, e.g. *jantarinho*, *animalinho*, and *nuinho*, derived from *jantar* 'dinner', *animal* 'animal', and *nu* 'naked', respectively. This type of formation is not productive in Brazilian Portuguese. A total of 15 missing items represent idiosyncratic formations, e.g. *frangainho* (from *frango* 'chicken'), *varginha* (from *vargem* 'floodplain'), *fontainha* (from *fontana* 'fountain'), *foicinho* (masculine diminutive from feminine noun *foice* 'sickle'), etc.

From the remaining 3 missing items, 2 are *bebezinha* and *bebezinhas*, singular and plural diminutive of feminine *bebé* 'baby', European Portuguese variant of Brazilian

Portuguese *bebê*. Since DGFL-ADJN only contains the homonym masculine forms *bebé* und *beb*ê, DIM2 only analyses the masculine diminutive forms. The other missing item is *rockzinhos*, from whose plural form DIM2 built the attested variant *rockezinhos*.

Just one of the 28 missing items from D-UP can be considered a true error of DIM2, namely *calças+N+DIM+F+PL*, whose lemma is the plurale tantum *calças* 'pants'. DIM2 only encodes *calça+N+DIM+F+PL* with the lemma in singular (which has a similar meaning in Portuguese). The remaining missing strings are lexical representations for which DGFL-ADJN provides no counterpart without the diminutive morpheme. These cases seem to be either errors in the original resources or irregular formations. For example, in DGFL-ADJN, *cebola+N+DIM+M+SG* maps to *cebolinho* 'chives'. The problem with this lexical representation is that there is no masculine gender noun *cebola* 'onion' in Portuguese, which is a feminine gender word. Regular diminutive formation is a gender preserving process in Portuguese. This prevents the derivation of a masculine gender noun such as *cebolinho* from a feminine gender noun. Accordingly, there is no lexical representation *cebola+N+M+SG* in DGFL-ADJN, only *cebola+N+F+SG*, as expected, and *cebolo+N+M+SG*, representing *cebolo* 'chives'.

In conclusion, due to the finite-state implementation of -*(z)inh*- diminutives compiled into DIM2, MorphoBr contains 60 times more diminutives than ADJN, the transducer DGFL-ADJN was compiled into. In morphological analysis, MorphoBr profits from a division of labour between DIM2 and ADJN, since it includes both: while the former only encodes regular, standard -*(z)inh*- diminutives (at least in Brazilian Portuguese), the latter encodes irregular and non-standard forms. Therefore, MorphoBr has a much wider coverage also in qualitative terms than the previous resources. This makes it far more suitable for analysing texts where both standard and non-standard formations are used.

5.4 Comparison to FreeLing's suffix rules

We compared the coverage of MorphoBr's converted verb-clitic clusters and diminutives generated by DIM2 against FreeLing 4.0 by checking if FreeLing recognizes the word form either because it is in its dictionary or because of one of its suffix rules (see section 2). To avoid any bias, all items without a lemma in FreeLing's dictionary were discarded. Since many diminutives are categorially ambiguous between noun and adjective, while sharing the same lemma, this category distinction was discarded and the resulting repetitions eliminated. This resulted in two test sets of unique pairs of word forms and lemmas: DIMINUTIVES, containing 415,098 diminutives, and V-CL-CLUSTERS with 893,796 verb clitic-clusters.

FreeLing missed 57.8% (240,074) of DIMINUTIVES, recognizing 1.3% (5,253) directly via dictionary lookup and 40.9% (169,771) via affixation rules. An inspection of the missed forms shows that FreeLing's rules do not handle the alternation between -*inh*- and -*zinh*- uniformly. While both *alegretinho* and *alegretezinho* are correctly lemmatized to *alegrete* 'planter', only *alegrinho* and *elefantinho* are lemmatized to *alegre* 'happy' and *elefante* 'elephant', respectively, but not the equally possible (and attested) variants *alegrezinho* and *elefantezinho*. The rules also fail to analyze plural forms like *florezinhas*, *luzezinhas*, etc. (more than 4500 similar cases).

FreeLing missed 0.5% (4,779) of V-CL-CLUSTERS, recognized five items directly via dictionary lookup and 99.5% (889,012) via affixation rules. A survey of the clusters FreeLing failed to analyse reveals that about half are not grammatical. In these clusters, clitic pronoun *nos* 'us' is attached to a verb form of the 2nd person plural of the future subjunctive tense, e.g. *zombarde-nos* (*zombar* 'mock'). According to Cunha and Cintra (1985, p. 399), this tense does not allow enclitics. FreeLing, however, did recognize thousands of these forms with other clitics, e.g. *zombarde-la*. These findings point to the need of revision of the treatment of clitic clusters by both MorphoBr and FreeLing. Notwithstanding these problems, we can conclude that current FreeLing's affixation rules for Portuguese are far more complete for clitics than for diminutives.

6 Conclusion

We have presented MorphoBr, a new wide-coverage full-form lexicon for Portuguese, released under a free, open source software license. It represents a two-fold contribution. First, previous freely-available resources were consolidated, removing several thousands of errors and gaps. Entries were converted to a uniform format using more mnemonic and linguistics-oriented tagging conventions. This format not only is more human-readable but also allows for straightforward compilation of finite-state morphological analysers.

MorphoBr, however, is not just a combination and correction of previous resources. Its main contribution is the systematic treatment of word-formation by computationally modeling the underlying linguistic regularities. As a test case of this approach, the formation of *-(z)inh-* diminutives was implemented in the paradigm of finite-state morphology. Previous resources either provide very incomplete lists of diminutives or formulate ad hoc rules that cover only a small part of the cases. By contrast, our systematic treatment of diminutive formation rules resulted in 170% more nouns and adjectives than DELAF-PB's and FreeLing's dictionaries combined. As regards the total amount of *-(z)inh-* diminutives, the finite-state implementation generated 60 times more such items than listed in these previous resources.

MorphoBr is still work in progress, but the experimental evaluation results seem promising. It clearly outperformed FreeLing's suffix rules in the coverage of *-(z)inh-* diminutives. This makes it more adequate for tasks dealing with texts where diminutives are very common.

Regarding the FreeLing coverage test, we are aware of the fact that many missing items could be avoided with improvements in FreeLing's affixation rules for Portuguese. Nevertheless, we leave as a future work testing whether this type of rules can deal as efficiently with all forms derived by our FST approach, which involves the formalization of intricate phenomena both at the morphotactic and morphographemic levels.

Other future related work includes: expanding our FSTs with other productive word-formation rules, also taking into account particularities of European Portuguese; reviewing the verb-clitic clusters and their annotation; dealing with MWEs; and implementing the grammatical word classes, e.g. determinants, conjunctions, etc.

The latter topic requires, however, a clarification about the lexical representations of these words, which, in turn, depends on the implementation of syntactic rules in a concrete grammatical formalism, GF and LFG in the case of our project. This means that the exact form of these entries can only be defined after developing grammar fragments in these formalisms covering the relevant grammatical phenomena.

Acknowledgements

We are indebted to Fabricio Chalub for his contributions to a previous version of Section 5.4. We also thank an anonymous reviewer for valuable comments and suggestions.

References

ALENCAR, L. F. de et al. JMorpher: A Finite-State Morphological Parser in Java for Android. In: BAPTISTA, J. et al. (Eds.). Computational Processing of the Portuguese Language. 11th International Conference, PROPOR 2014. São Carlos/SP, Brazil, October 6-8, 2014. *Proceedings*... Heidelberg: Springer, 2014, p. 59-69.

BAZENGA, A. M. Sufixos avaliativos -*inh-/-zinh*- em português: da morfologia à pragmática da ironia verbal. *Pensardiverso*, Funchal, v. 3, p. 115-130, 2012. Disponível em: <https://digituma.uma.pt/handle/10400.13/1729>. Acesso em: 2 abr. 2018.

BEESLEY, K. R.; KARTTUNEN, L. *Finite state morphology*. Stanford, California: CSLI, 2003.

BICK, E. PALAVRAS: a constraint grammar-based parsing system for portuguese. In: SARDINHA, T. B.; FERREIRA, T. L. S. B. (Org.). *Working with Portuguese Corpora*. [S.l.]: Bloomsbury Academic, 2014. p. 279-302.

BUTT, M. et al. *A Grammar Writer's Cookbook*. Stanford, California: CSLI, 1999.

CIPRO NETO, P. Está errado dizer “cheinho" e “sainha"? In: CIPRO NETO, P. *Dicas do Pasquale*. [S.n.t.]. Disponível em: <http://www.educacional.com.br/espacopasquale/dicas.asp?intPagAtual=10&>. Acesso em: 11 mai. 2018.

COSTA, J. A.; MELO, S. *Dicionário da Língua Portuguesa*. 6. ed. Porto: Porto Editora, 1991.

CUNHA, C.; CINTRA, L. F. L. *Nova gramática do português contemporâneo*. Rio de Janeiro: Nova Fronteira, 1985.

DIPPER, S. *Implementing and documenting large-scale grammars – German LFG*. 2003. 359 f. Tese (Doutorado) - Philosophisch-Historische Fakultät, Universidtät Stuttgart, 2003.

ELEUTÉRIO, S. et al. A system of electronic dictionaries of portuguese. *Lingvisticae Investigationes*, v. 19, n. 1, p. 57-82, 1995. Disponível em:

<http://label.ist.utl.pt/publications/docs/Eleuterio_et_al_95.pdf>. Acesso em: 18 set. 2018.

FALK, Y. *Lexical-Functional Grammar*: an introduction to parallel constraint-based syntax. Stanford, California: CSLI, 2001.

FRADIN, B. *Abbréviation des gloses morphologiques*. Paris: Laboratoire de Linguistique Formelle, Université Paris-Diderot, [s.d.]. Disponível em: <http://www.llf.cnrs.fr/fr/node/60>. Acesso em: 16 set. 2018.

GARCIA, M.; GAMALLO, P. Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática*, Braga, v. 2, n. 2, p. 59-67, 2010. Disponível em: <http://linguamatica.com/index.php/linguamatica/article/view/56>. Acesso em: 15 out. 2018.

GARCIA, M. et al. PoS-tagging the Web in Portuguese: national varieties, text typologies and spelling systems. *Procesamiento del Lenguaje Natural*, v. 53, p. 95-101, 2014. Disponível em: <http://www.taln.upf.edu/pages/sepln2014/full_papers/edited_paper_21.pdf>. Acesso em: 18 set. 2018.

HULDEN, M. Foma: a finite-state compiler and library. In: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 12, 2009, Athens. *Proceedings*... [S.l.]: Association for Computational Linguistics, 2009. p. 29-32. Disponível em: <http://www.aclweb.org/anthology/E09-2008>. Acesso em: 18 jun. 2018.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. London: Pearson, 2009.

LAPA, M. R. *Estilística da língua portuguesa*. São Paulo: Martins Fontes, 1982.

LEACH, G.; WILSON, A. *Recommendations for the morphosyntactic annotation of corpora*. [S.n.t.], 1996. Disponível em: <http://www.ilc.cnr.it/EAGLES/pub/eagles/corpora/annotate.ps.gz>. Acesso em: 15 set. 2018.

LIMA, R. *Gramática normativa da língua portuguesa*. 49. ed. Rio de Janeiro: José Olympio, 2011.

MONTEIRO, J. L. *Morfologia portuguesa*. 2. ed. Fortaleza: EDUFC, 1987.

MCDONALD, R. et al. Universal dependency annotation for multilingual parsing. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 51, 2013, Sofia. *Proceedings*... [S.n.t.], 2013. p. 92-97. Disponível em: <https://www.aclweb.org/anthology/P13-2017>. Acesso em: 18 set. 2018.

MUNIZ, M. C. M. *A construção de recursos lingüístico-computacionais para o português do Brasil*: o projeto Unitex-PB. 2004. 92 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

NIVRE, J. et al. *Universal Dependencies 2.1*. Prague: Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2017 Disponível em: <http://hdl.handle.net/11234/1-2515>. Acesso em: 18 set. 2018.

NOGUEIRA, S. As luzinhas ou as luzezinhas de Natal? In: NOGUEIRA, S. *Dicas de português*: Temas polêmicos. [S.l.]: Globo, 2010. Disponível em: <http://g1.globo.com/educacao/blog/dicas-de-portugues/post/temas-polemicos-3.html>. Acesso em: 15 mai. 2018.

PADRÓ, L.; STANILOVSKY, E. Freeling 3.0: Towards wider multilinguality. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 8, 2012, Istambul. *Proceedings*... [S.n.t.], 2012. p. 2473-2479. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf>. Acesso em: 15 set. 2018.

POLLARD, C.; SAG, I. A. *Head-driven phrase structure grammar*. Stanford: CSLI, 1994.

RANCHHOD, E.; MOTA, C.; BAPTISTA, J. A computational lexicon of Portuguese for automatic text parsing. In: STANDARDIZING LEXICAL RESOURCES, 1999, College Park, Maryland. *Proceedings…* [S.n.t], 1999. p. 74-80. Disponível em: <http://www.aclweb.org/anthology/W99-0511>. Acesso em: 15 set. 2018.

RANTA, A. *Grammatical Framework*: programming with multilingual grammars. Stanford, California: CSLI, 2011.

RIO-TORTO, G. Formação de avaliativos. In: RIO-TORTO, G. et al. (Org.). *Gramática derivacional do português*. Coimbra: Coimbra University Press, 2016. p. 357-389.

ROCHA, L. C. de A. *Estruturas morfológicas do português*. 2. ed. São Paulo: Martins Fontes, 2008.

VILLALVA, A.; SILVESTRE, J. P. *Introdução ao estudo do léxico*: descrição e análise do português. Petrópolis: Vozes, 2014.

ŠEVČÍKOVÁ, M. Modelling morphographemic alternations in derivation of Czech. *The Prague Bulletin of Mathematical Linguistics*, Prague, v. 110, p. 7-42, 2018. Disponível em: <https://ufal.mff.cuni.cz/pbml/110/art-sevcikova.pdf>. Acesso em: 4 mai. 2018.

Recebido em dia 05 de outubro de 2018.
Aprovado em dia 12 de novembro de 2018.