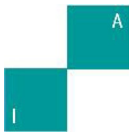
Inteligencia Artificial 25(69) (54), 87-106
doi: 10.4114/intartif.vol25iss69pp87-106



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Detection of Loanwords in Angolan Portuguese: A Text Mining Approach

Timóteo Sumbula Muhongo[1,2] Pavel B. Brazdil [1,3], Fátima Silva[2]

[1] LIAAD - INESC TEC, Porto, Portugal
[2] CLUP - FLUP, University of Porto, Porto, Portugal
[3] FEP, U. Porto, Porto, Portugal
Email: timuhongo@hotmail.com, pbrazdil@inesctec.pt, mhenri@letras.up.pt

Abstract. Angola is characterized by many different languages and social, cultural and political realities, which had a marked effect on Angolan Portuguese (AP). Consequently, AP is characterized by diatopic variation. One of the marked effects is the loanwords imported from other Angolan languages. Our objective is to analyze different Angolan texts, analyze the lexical forms used and conduct a comparative study with European Portuguese, aiming at identifying the possible loanwords in Angolan Portuguese. This process was automated, as well as the identification of all loanwords' cotexts. In addition, we determine the lexical class of each loanword and the Angolan language of its origin. Most lexical loanwords come from the Kimbundu, although AP includes loanwords from some other Angolan languages too. Our study serves as a basis for preparing an Angolan regionalism dictionary. We noticed that more than 700 identified loanwords do not figure in the existing dictionaries.

Keywords: Text Mining, Natural Language Processing, Incremental Text Processing, Contrastive Lexicology, Neology.

1 Introduction

This research aims at a comparative study of lexical forms in Angolan Portuguese (AP) and European Portuguese (EP) to identify the specificities of Portuguese in Angola, including some lexical loanwords from Angolan languages that occur in the writings portraying the current Angolan linguistic reality.

The motivation for our investigation is twofold. On the one hand, we consider that there is a gap in the knowledge of certain aspects of AP, particularly in relation to lexical loanwords, their meaning, and their usage, which can be captured in dictionaries. On the other hand, there is also a need to create a dictionary of Angolan regionalisms that can help students and teachers in Angola and those interested in understanding the Literature and History of Angola.

ISSN: 1988-3064(on-line)

We advocate that the lexical forms reflect historical and cultural realities and, as an invaluable heritage, deserve attention and protection. Assuming that language, culture and history form a special trinomial in lexematics, we can relate it in the description and understanding of the functioning of lexical forms of Angolan and European Portuguese.

This paper consists of five sections. In section 1, *Introduction*, we present the object of the study, the purpose, and the motivation. In section 2, *Related Work*, we focus on the relationship and difference between other works in the fields of lexematics and computational linguistics on this subject and our work. In Section 3, *Methodology*, we describe the steps used to achieve our goal. We present the corpus of lexicon extraction of Angolan and European Portuguese, and briefly describe it, considering specific classification parameters. Then we present the methodology followed in our research. It discusses some NLP and text mining techniques used in this process. In Section 4, we analyze the candidates for lexical loans, by identifying the elements that appear in AP and not in EP. We detect the loanwords, determine their etymology and syntactic category. Besides, we detect the loanwords cotext and present the prototype of the dictionary of Angolan regionalisms. In Section 5, *Conclusions and Future Work,* we refer to the need to extend this study to some nominal and multi-word units and present the dictionary of Angolan regionalisms prototype.

Figures should be centered and included in the text. You may find an example in figure 1. Tables should follow a similar scheme. You may find an example in table 1.

• 2 Related Work

In this section, we present some studies that deal with loanwords from Angolan languages and studies in the area of computational linguistics, which are related to our subject. We proceed to explain the dissimilarity between the studies developed and our scrutiny.

2.1. Area of lexematics

There are not many studies in this area, in particular to computational approaches. There has been more interest in this area recently, which has been reflected in the emergence of studies such as those by Costa [1] and Silva [2]. The first is limited to Umbundu and the second to Portuguese and Kimbundu, in an attempt to find equivalents in the field of health. Neither Costa [1] nor Silva [2] used computational approaches in their analysis and did not consider Named Entity Recognition (NER), which was used by Maurice Gross [3].

Lino and Dechamps [4] made a comparative study of legal terminology of European Portuguese and French. For this purpose, they used the hyperbase, a program for semi-automatic treatment of corpora. Muhongo [5] also discussed lexical loans from Angola and used the AntConc and Concapp programs for semi-automatic processing of the corpus. Unlike these works, we use a computational approach to identify Angolan lexical loanwords by lexical class, so far regardless of the Angolan language from which they came from.

2.2. Area of Computational Linguistics

In this section, we consider the research in computational linguistics, which is relevant for our objective and the methodology adopted.

Canosa *et al.* [6] describe a method to build a tool designed to recognise named geographical entities in medieval texts. However, the new tool was developed from the contemporary language modules of LinguaKit, a natural language processing toolkit, with which gazetteers, a list of medieval toponyms, was developed. They observed patterns for the improvement and implementation of new rules for the recognition of geographical names. After the list of geographical entities, the contextual triggers were the determining resource improvement recall.

Taking the performance evaluation of predictive models as a starting point, Pinto *et al*. [7] evaluated several natural language processing tools with their default configuration, while performing a set of standard tasks (e.g. tokenisation, POS tagging, chunking, and NER) on popular datasets covering newspapers and social networks, using programming in Java and Python. Precision, recall, F-measure, micro and macro averages were used as performance measures of the automated method. The results obtained were valuable to narrow down its choice of natural language processing toolkit.

Gamallo and Garcia [8] propose a feature-based Named Entity Classification (NEC) system that combines named entity extraction with simple language-independent heuristics. They automatically extracted a geographic dictionary (gazetteers) of named entities, using semi-structured information from Wikipedia, such as infoboxes and classification trees. Language-independent heuristics were used to disambiguate and classify entities that were already recognized in the text. Furthermore, they compared the performance of the feature-based system with that of a supervised NEC module implemented for FreeLing. Experiments were carried out on Portuguese text corpora considering several domains and genres.

Starting from the analysis of theses and dissertations abstracts, Iriguti and Feltrim [9] sought to automatically identify sentence-level categories that make up rhetorical structures of scientific abstracts. Their goal was to evaluate the impact of different sets of attributes on the implementation of rhetorical classifiers for scientific abstracts written in Portuguese. For this, surface attributes are extracted as TF-IDF values and selected with the x^2 test, morphosyntactic attributes implemented by the AZPort classifier, and attributes extracted from word embedding models.

• 3 Methodology

The methodology adopted can be divided into two parts. The first part includes the following steps.

1. Build corpora of AP and EP;
2. Carry out preprocessing;
3. Extract candidate loanwords, which are lexical elements of AP that do not appear in EP;
4. Analyze (manually) the loanwords candidates to identify the lexical loanwords;
5. Automatically extract the cotexts and frequencies of the loanwords.

More details on these steps are given in various subsections of this section. In addition, Fig.1provides an an overview of the method.

The second part of the methodology involves the following two steps:

1. Describe the loanwords formation processes;
2. Propose a dictionary of Angolan regionalisms.

More details on both issues are given in Section 4.

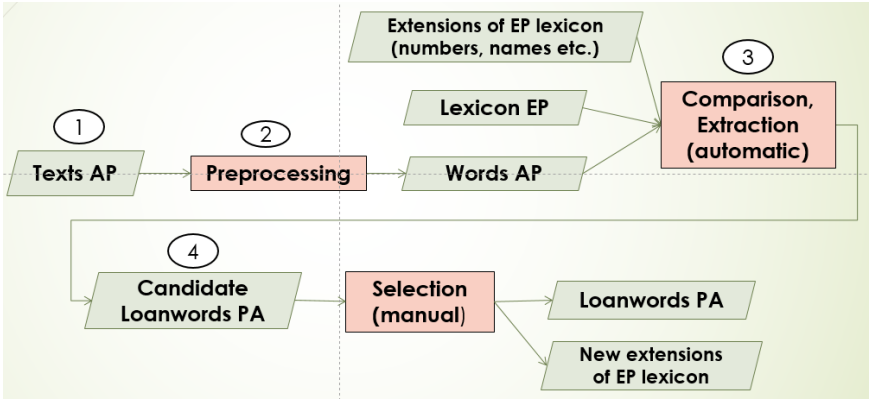


Fig.1. Extraction of loanwords

3.1. Constitution of the corpus

Our corpus of AP extraction includes:

- *A Conjura* (2008), which has 10,392 tokens;
- *Jornal de Angola* (2019-2020), composed of twenty-three documents, totaling 82,572 tokens;
- *Telejornal TPA* (2020), composed of five documents, which has 30,524 tokens.

The total number of tokens in the three corpora above is 123,488. The third corpus listed, i.e., *Telejornal TPA* (2020), was obtained as follows. First, we have recorded *Telejornal TPA* (2020) using the *Audacity* system. The automatic transcription was done with the *Dictate* program. Following Torruella & Llisterri [10], Adam [11], Sinclair [12], and Llamazares [13] corpora classification proposals, we consider that our corpus features are being both in written and oral form, monolingual, general, closed, synchronous, textual and coded...

Regarding EP, we have used the lexicon for Portuguese provided by Pablo Gamallo of the University of Santiago de Compostela, which contains 1,110,724 tokens. Each entry includes an inflexed form (e.g., *descarregámos* – *'downloaded'*), the corresponding lemmatized form (e.g., *descarregar* – *'download'*) and information about lexical class and other information (e.g., *VMIS1P0*).

3.2. Preprocessing and Identifying Words of Given Lexical Class

Preprocessing normally involves tokenization, removal of punctuation, removal of numbers, conversion to lower case, removal of stop words and lemmatization [14]. Further operations involve identification of words of a certain lexical class.

In this work, we have not followed this rather traditional path, but instead followed a more convenient solution, which involved the use of package *udpipe* of R [15]. This package can process the input text, carry out tokenization, and, for each token, determine its lemmatized form and lexical class (upos).

In this study, we are interested in four lexical classes: *verbs*, *nouns, adjectives, and proper nouns*, as the loanwords in AP belong predominantly to these four classes. Fig. 2 shows the R code used to identify words and their respective lexical class (upos) from "CONJURA.txt".

```
Sys.setlocale("LC_ALL","pt_BR.UTF-8")
text.conjura <- readLines("CONJURA.txt")
install.packages("udpipe")
library("udpipe")
udmodel <- udpipe_download_model(language = "portuguese")
udmodel <- udpipe_load_model(file = udmodel$file_model)
text.annot.conjura <- udpipe_annotate(udmodel, x = text.conjura)
text.annot.conjura <- as.data.frame(text.annot.conjura, detailed = TRUE)
text.annot.conjura  <-  text.annot.conjura[,c("token_id","token","lemma",
"upos")]
```

Fig.2. Instructions of R and *udpipe* package used to identify tokens and their lexical class

Table 1. shows a few examples of tokens identified, together with their lemmas and lexical class.

Table 1 – Some tokens and their lexical class

token_id	token	lemma	upos
1	a	o	DET
2	Conjura	Conjura	PROPN
1	em	em	ADP
2	memória	memória	NOUN
3	de	de	ADP
4	Pedro	Pedro	PROPN
5	conta	contar	VERB

This way the obtained information can be used to extract lemmas of a particular lexical class. The following instructions show how we can extract, for instance, verbal lemmas.

```
index.verbs <- text.annot.conjura$upos == "VERB"
verbs.conjura <- text.annot.conjura[index.verbs, "lemma"]
```

This process can be repeated for the other lexical classes of interest. Table 2 shows the number of occurrences of the words identified in our AP corpus for given lexical classes.

Table 2 – Numbers of occurrences of lemmas for the four lexical classes

	Verb	Noun	Adj	PropN
A Conjura	5.072	8.442	2.321	2.496
Jornal de Angola	70.425	175.811	43.534	160.104
Telejornal	3.456	7.523	1.890	1.434
Total	**78.953**	**191.776**	**47.745**	**163.994**

Note that a particular word may appear several times in a given text. It is, of course, not necessary to analyze the repeated occurrences but consider just one of its occurrences. Table 3 shows the number of occurrences after eliminating repetitions.

Table 3 – Numbers of non-repeated occurrences of lemmas for the four lexical classes

	Verb	Noun	Adj	PropN
A Conjura	1.235	2.960	1.008	530
Jornal de Angola	2.029	12.295	6.277	26.433
Telejornal	540	1.855	588	389
Total	**3.804**	**17.110**	**7.873**	**27.352**

We realize that the number of elements to analyze has been substantially reduced. In fact, for some lexical classes (e.g., verbs and nouns), the numbers have been reduced by less than 10%.

Each group of lexical items (e.g., verbs, etc.) may contain some loanwords. However, analyzing several thousands of cases manually is laborious. To further reduce the manual effort, we use a computational method that involves a comparison with a lexicon of EP. The details of this method are described in the next subsection.

3.3. Identify candidates of lexical loanwords

Different types of loanwords were identified in literature on neology, as Fig. 3 shows. Here we focus on interlinguistic loanwords.

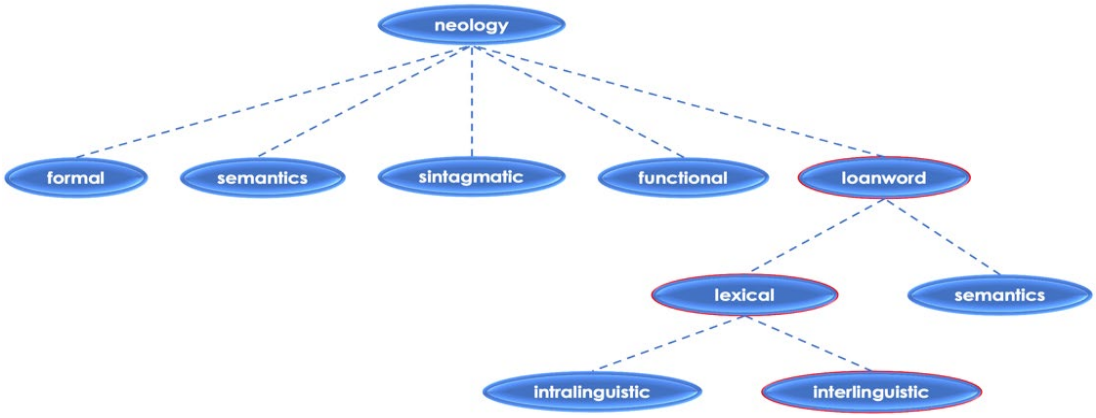


Fig. 3. Types of neology [16, 17].

The method used involves a comparison of lexical elements of AP of a certain lexical class (e.g., verbs) with lexical elements that appear in a given lexicon EP. This is done with the instruction

setdiff(lex.elem.AP, lex.elem.EP),

which identifies the subset of lexical elements of AP (lex.elem.AP) that do not appear among the lexical elements of EP (lex.elem.EP). Some candidate loanwords that were extracted are shown in Fig. 4.

Fig. 4. Some extracted candidate loanwords

Table 4. presents the number of candidate loanwords for AP that were extracted this way from the corpus. They are organized by lexical class.

Table 4 – Number of loanword candidates by lexical class

	Verb	Noun	Adj	PropN	Total
A Conjura	92	234	68	227	621
Jornal de Angola	1.407	2.112	1.908	12.142	17.569
Telejornal	53	91	32	261	437
Total	**1.552**	**2.437**	**2.008**	**12.630**	**18.627**

As the candidate loanwords include various words that are not loanwords, it was necessary to examine all these candidates to identify the true loanwords. This topic is described in the next subsection.

3.4. Identify the lexical loanwords among the candidates

In this step we need to separate the true loanwords from others that were erroneously returned as candidate loanwords. These incorrect candidates are of several types:

1. Numbers (e.g., three, ten);
2. Personal names (e.g., José) (the lexicon of EP does not include personal names);
3. Other words from EP that are not in the lexicon of EP (e.g., inusitar);

4. Words that have been misspelled (e.g., aconselher, afirmer);
5. Specific words (e.g., imberber, juditar).

Groups 1 and 2 above need to be considered, because the used lexicon of EP does not include these elements. Consequently, it is necessary to analyze the candidates and identify true loanwords. Moreover, it is also useful to store the examples of the other groups, as we shall see further on.

3.4.1 Identifying Verbal Loanwords

The list of loanword candidates that are verbs includes both true and false loanwords. Consequently, it is necessary to analyze each candidate to determine to which group it belongs. This analysis was done manually.

Regarding the loanwords in *A Conjura*, we analysed the 92 candidates referred to earlier (see Table 4) and extracted 12 true loanwords, which are listed in Table 5, together with their frequency. The elements that include "+" in the column "Dic" are listed in the *Dictionary of Angolan Regionalisms* (Ribas, 2014). Those with a negative dash (-) are not listed there, nor in the dictionary or in the *Dicionário da Língua Portuguesa Contemporânea* (*Academia das Ciências de Lisboa & Fundação Calouste Gulbenkian, 2001).*

Table 5 – Verb loanwords in *A Conjura*

N.º	Verb loanword	Freq.	Dic.	N.º	Verb loanword	Freq.	Dic.
1	baçular	1	+	**7**	molumbar	1	+
2	coxilar	2	+	**8**	muturir	1	+
3	facar	1	–	**9**	ngar	1	–
4	kuatar	1	+	**10**	sunguilar	1	+
5	malembelember	1	–	**11**	uandi	1	–
6	massembar	1	+	**12**	xuaxulhar	2	–

Some false verbal loanwords are shown in Table 6. They include words that are not in our EP lexicon (e.g., cinquentar), so called *specific verbs* that were introduced by the author (e.g., imberber), misspelled words (e.g., aconselher),and misclassified words (e.g., dande, diquixi). The last group includes words of the nominal class but were classified as being of the verbal class.

Table 6 – False verbal loanwords in *A Conjura*

N.º	Not in EP lexicon	Specific	Misspelled	Misclassified
1	cinquentar	imberber	aconselher	dande
2	desconseguir	juditar	afirmer	diquixi
3	inusitar	preconceituar	cabar	kunene-bu
4	magistrar	charar	concordavar	muhatu
5	malcasar	–	consumirar	n'dalatando
6	oitentar	–	descobrar	–

7	safrar	–	...	–
Total	7	4	64	5

The process of eliding elements from the candidate list was repeated with the data from *Jornal de Angola* and *Telejornal*. Table 7 shows the result of this procedure. The row labelled "Verb Dic-" indicates how many verbal loanwords do not appear in the existing dictionaries.

Table 7 – Quantification of verbal loanwords

	A Conjura	**Jornal de Angola**	**Telejornal**	**Total**
Verb	12	52	2	66
Verb Dic. -	5	26	2	33

3.4.2 Identifying nominal loanwords

A similar process was used to identify the nominal loanwords. The nominal loanwords identified from *A Conjura* are shown in Table 8. Table 9 shows the loanwords identified from the three source texts used.

Table 8 – Nominal loanwords from *A Conjura*

N.º	**Noun loanword**	**Freq.**	**Dic.**	**N.º**	**Noun loanword**	**Freq.**	**Dic.**
1	andua	1	+	**10**	macala	1	+
2	calunga-ya-meia	1	–	**11**	machila	3	+
3	cazuela	1	–	**12**	macololo	2	–
4	dicamba-dia-ngalafa	1	–	**13**	maka	4	+
5	dicanza	1	+	**14**	monangamba	1	+
6	diquixi	2	+	**15**	mujimbu	1	+
7	gindungo	1	+	**16**	quilamba	1	+
8	kissangua	1	+	**17**	quilumba	1	+
9	libata	1	+	**18**	quindumba	3	+

This procedure of elision of elements from the candidate list was repeated with the data from *Jornal de Angola* and *Telejornal*. Table 9 shows the result of this procedure.

Table 9 – Number of nominal loanwords identified

	A Conjura	**Jornal de Angola**	**Telejornal**	**Total**
Noun	101	128	15	244
Noun Dic.-	30	112	–	142

3.4.3 Identifying adjectival loanwords

Having obtained 68 candidate loanwords with the automatic method described in Section 3.3, it was necessary to analyze all the cases manually to separate false loanwords from true ones. The text *A Conjura* led to 12

adjective loanwords, which can be seen in Table 10, together with their respective frequency and whether they appear in the existing dictionaries.

Table 10 – Adjectival loanwords from *A Conjura*

N.º	Adj. loanword	Freq.	Dic.	N.º	Adj. loanword	Freq.	Dic.
1	agindungado	1	+	**7**	ngo	1	+
2	ambaquense	1	–	**8**	nzua	1	+
3	andembo-ya-tata	2	–	**9**	quindumbo	1	–
4	cuamato	9	+	**10**	sapalalo	1	–
5	diculu	1	+	**11**	tchibita	1	–
6	muxito	1	+	**12**	uanga	1	+

The process of elision of elements from the candidate list was repeated with the data from *Jornal de Angola* and *Telejornal*. Table 11 shows the resulting adjectival loanwords.

Table 11 – Number of adjectival loanwords identified

	A Conjura	Jornal de Angola	Telejornal	Total
Adj	12	39	1	52
Adj Dic. -	5	31	–	36

3.4.4 Identifying loanwords that are proper nouns

Having processed the text *A Conjura,* we have identified 227 candidate loanwords. They were analyzed manually, resulting in 68 true loanwords, all of which are proper nouns. Some examples are shown in Table 12.

Table 12 – Proper noun loanwords from *A Conjura*

N.º	PropN loanword	Freq.	Dic.	N.º	PropN loanword	Freq.	Dic.
1	Amboim	1	+	**7**	Massangano	2	–
2	Bungo	7	+	**8**	Pacavira	1	–
3	Cambambe	2	–	**9**	Pungo Andongo	4	+
4	Humbe	8	+	**10**	Quissama	3	+
5	Kuanhama	1	+	**11**	Quissongo	4	+
6	Magombala	7	–	**12**	Zenza	1	–

The procedure of identification of true loanwords was repeated for *Jornal de Angola* and *Telejornal*. Table 13 shows the proper noun loanwords resulting from this procedure.

Table 13 – Number of proper noun loanwords identified

	A Conjura	Jornal de Angola	Telejornal	Total
PropN	68	1.286	68	1.422
PropN Dic -	20	666	26	712

3.4.5 Summary of the identified loanwords

After the extraction of lexical loanwords, described in the previous subsections, we verified that *Jornal de Angola* led to more loanwords for all lexical classes, as can be seen in Table 14. In summary, 1.784 loanwords were identified, 923 of them not appearing within the above-mentioned dictionaries.

Table 14 – Numbers of loanwords by lexical class

	Verb	Noun	Adj	PropN	Total
A Conjura	12	101	12	68	193
Jornal de Angola	52	128	39	1.286	1.505
Telejornal	2	15	1	68	86
Total	**66**	**244**	**52**	**1.422**	**1.784**

- **3.5. Extracting the cotext of loanwords**

Our subsequent objective was to extract the cotext of loanwords to analyze how they are used. For some loanwords, it is possible to extract the preceding word, which is in some functional relation to the loanword. In general, it is necessary to use dependency relations among lexical items to extract the cotext. Let us analyze an example involving the loanword *maka* (meaning *some kind of problem*), which appears in the sentence "*O Silva tinha que arranjar uma maka* (Silva had to arrange a problem)". Fig. 5 shows the output of dependency parsing carried out by *udpipe* package of R[1]. As can be seen, the loanword "*maka*" is the object of the verb "*arranjar*". This suggests the cotext, in this case, is "*arranjar uma maka*".

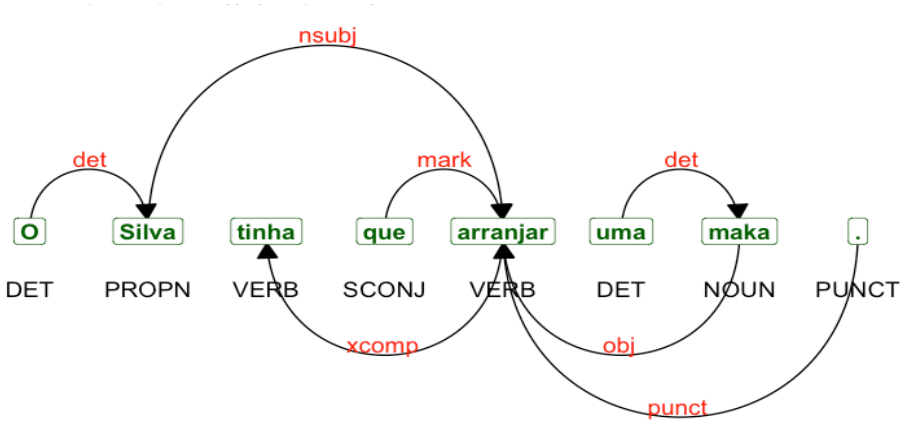


Fig. 5. Dependency relations for *maka* loanword obtained with "udpipe" package

[1] Various tools exist in different programming languages that can be used for dependency parsing. One example is, for instance, spaCy of Python.

Table 15 presents examples of several loanwords and cotexts that were identified this way. In it, we present the loanwords, etymology, meaning, document from which the loanword was taken, cotext, and equivalence. This work has resulted in 465 entries spanning approximately 80 pages.

Table 15 – Examples of some loanwords and their cotext

Loanword	Etymology	Meaning	Doc	Phrase	Cotext	Meaning in English
buxila	kimbundu	*n. m.* & *f.* **1.** Son or daughter of slaves. **2.** Son or daughter of a slave girl with a free man, but born in the house in which she serves. **3.** Criollo.	1	1381	paixão por uma **buxila**	passion for a **slave's daughter**
				1428	tímido e a **buxila**	shy and the **slave's daughter**
				1437	cólera contra a inocente **buxila**	cholera against innocent **slave's daughter**
				1828	embarque da **buxila**	boarding of the **slave's daughter**
machila	kimbundu	*n. f.* **1.** Palanquim. **2.** A chair with a top and curtains, which was suspended from a bamboo pole and carried on the shoulders of two men.	1	167	subido em sua **machila**	climbing in the **king's chair**
				239	carregadores de **machila**	**king's chair** porters
				244	carregadores de trabalhadores **machila**	bearers of the **king's chair**
maka	kimbundu	**1.** *n. f.*Conversation; subject; novelty; discord; litigation; conflict; racket. **2.** *Bras.* Incolumanca. **3.** *adj.* Problematic.	1	125	ouvirem as **makas**	listen to the **problems**
				565	arranjar uma **maka**	fixing a **problem**
				1267	reconstruir alguma **maka**	rebuild a **problem**
				1590	centro das **makas**	**problematic** person
muxima *	kimbundu	*n. m.* & *f.* **1.** A person of good character. **2.** Benefactor. *n. m.* **3.** Heart. **4.** *Fig.* To do something sincerely. **5.***v.* To captivate, flatter.	1	402	briguento, mas de boa **muxima**	quarrelsome, but of good **character**
				776	inquieta **muxima**	restless **heart**
quindumba	kimbundu	*n. f.* **1.** Sparing. **2.** Raising of hair or feathers.	1	373	olho ao cheiroso corpo **quindumba**	attention to the smelling body of **pulp**
				463	corpos e as **quindumbas**	body and **hair-raising**
				1201	perfumada **quindumba**	scented **hair**

- **3.6 Advantages of the automatic process**

  - As was pointed out in Section 3.1, the three corpora used in our study include 123,488 tokens. Scanning this text manually with the objective of identifying all the loanwords (1,784 in total) would represent a major effort. The method described in the previous sections helps to reduce this effort substantially. Automatic procedures can be used to reduce the original set of 123,488 tokens to 18,627 candidate loanwords. This represents a reduction to about 15% of the original set.

  - Further gains can be obtained by incorporating a method that can recognize some false loanwords. In Section 3.4., we have shown that false loanwords are of different types, such as numbers, personal names, words not included in the existing EP lexicon, misspelled words, and specific words. So, it is possible to use an existing detector (or build a new one) for each of

these types. The simplest kind of detector uses just a list of words of a specific type (e.g. personal names). We plan to use such detectors in our future work.

4. Analysis of loanwords

In order to analyze, organize and visualize the information regarding the lexical loanwords, in this section we discuss the following issues:

- Determining the etymology of loanwords;
- Determining the syntactic category of loanwords;
- Prototype dictionary of Angolan regionalisms.

More details about each of these issues are given in the following subsections.

- **4.1 Determination of loanword etymology**

To determine the etymology of loanwords, we used the *Dictionary of Angolan Regionalisms* [18], and the *Dictionary of Contemporary Portuguese* (2001), which have helped to identify the etymology of the lexemes. Among the 1.784 lexical loanwords of Angolan language identified, most (872) are from Kimbundu, and 380 are from Umbundu, although many other languages are also represented, as can be seen in Table 16.

Table 16 – Quantification of lexical loanwords by language

		A Conjura	J. de Angola	Telejornal	Total	Language
Etymology	Kimbundu	146	694	32	872	Bantu
	Umbundu	12	354	14	380	Bantu
	Kikongo	3	117	6	126	Bantu
	Cokwe	3	115	4	122	Bantu
	Ngangela	3	31	–	34	Bantu
	Kwanyama	4	18	2	24	Bantu
	Nyaneka	4	12	2	18	Bantu
	Herero	1	3	–	4	Bantu
	Koisan	–	1	–	1	~Bantu
	Ngoyo	–	1	–	1	Bantu
	Lingala	–	1	–	1	Bantu
	Português em Angola	7	109	18	134	Latina
	Termo Regional	4	2	–	6	Bantu
	Total	**187**	**1.458**	**78**	**1.723**	**3**

Fig. 6 provides a visualization in the form of a bar chart of the data shown in Table 16.

4.2. Categorization of loanwords and their formation process

In this section, we discuss some ideas on the formation process of loanwords. As Pruvost & Sablayrolles [30] and Caldas [16] suggest, some loanwords are

- morphological compounds (*hidroluachimo, hidrochicapa*),
- morphosyntactic (*malembelember, matabicho, diculundundu*),
- syntagmatic (*dicamba-dia-ngalafa, calungaya-meia, suco-yo-bába, muezezambi-ya-mema*) [16, 17].

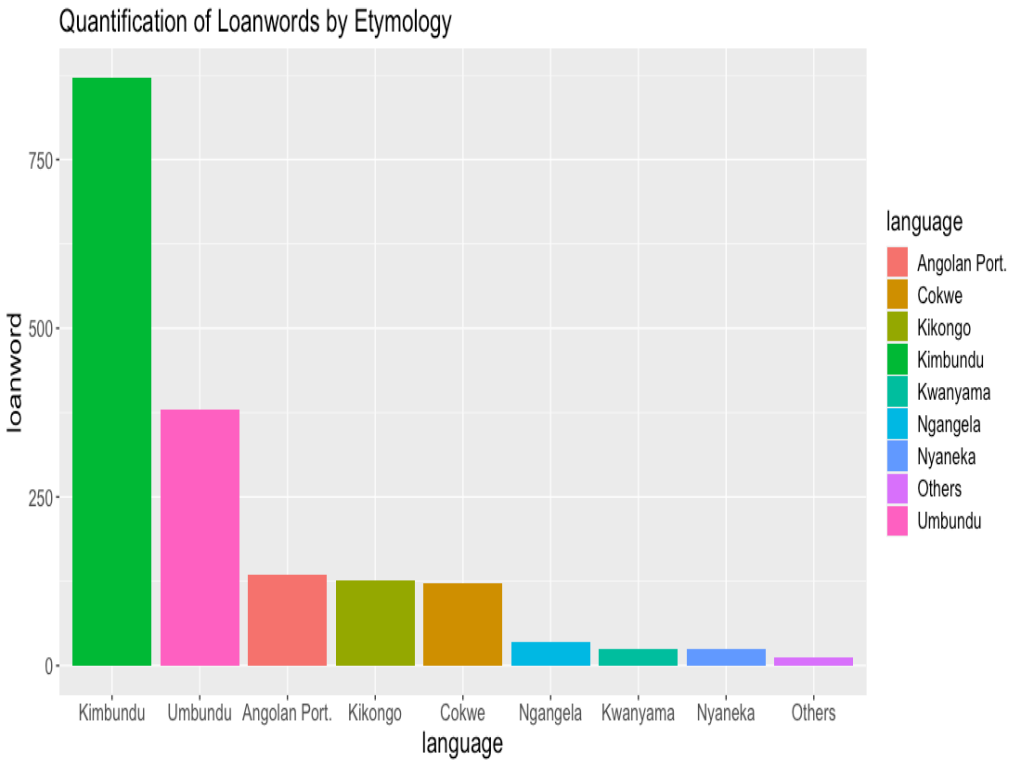


Fig 6. Quantification of lexical loanwords by etymology

Others loanwords are formed by:

- affix (*quimbandice, agindungado, Ambriz, Amboim*) [19],
- allomorphy (*jindungo/gindungo, imbondeiro/embondeiro, funji/fungi*),
- truncation (*calu, sobos, quiamba, mutu*),
- conversion (*muxima*) [20],
- vocabular crossing (*muxiluanda, monangamba, ingombota, chinangol, musangola, mundele, refriango, Textang, bessangana*) [21, 22].

We notice that some of the loanwords were formed by phonetic processes, such as:

- velarization (tchilengue > quilengue),
- denationalization (*Ndembo > Dembo; Ndandji > Dande; mbombo > bombô; Mbengu > Bengo; ngoma > goma; Kambinda > Cabinda*),
- prosthesis (*Mbaka > Ambaca; Ngola > Angola; nguba > ginguba*),

- apheresis (*kutunga > tunga; Elunda > Lunda; Ndembo > Dembo; Ndandji > Dande; mbombo > bombô; Mbengu > Bengo; ngoma > goma*),
- syncope (*Katombelwa > Catumbela*).

In addition, other loanwords present

- semantic extension (e.g. muxima, capeto).

- **4.4. Prototype dictionary of Angolan regionalisms**

In this section we present the prototype of the dictionary that we will develop in digital format. The survey of the units to be included in the dictionary is largely the result obtained, as is the application of specific lexicographical techniques, and validation by an expert team. In addition to explaining the criteria underlying the proposal of this prototype dictionary, and more specifically, a model of a lexicographical file, we propose to analyze a set of entries that may serve as a basis for building that dictionary, which constitutes a project to be developed following the approach in [23]. Our study Fig. 7 shows the regions where the respective language is spoken.

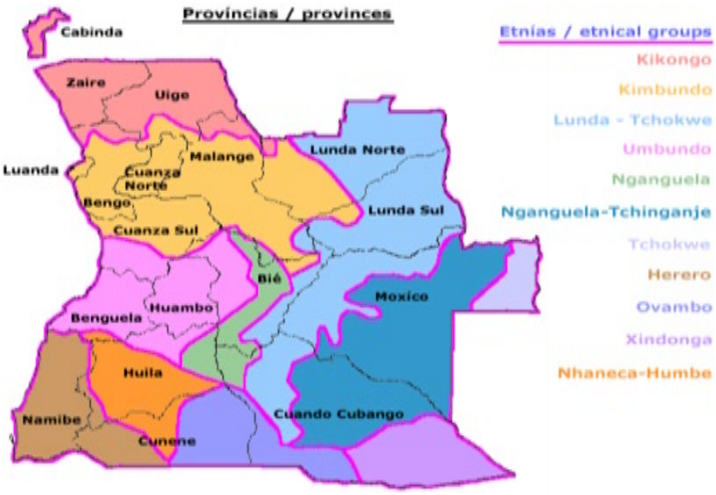


Fig. 7. People and languages of Angola, [24]

The collected data will be organized and presented in a lexicographical file, creating a prototype of the dictionary that we will develop in electronic format. Table 15, shown before, has just the basic information. It can be compared to a dictionary that can be consulted by users. Fig 8 includes additional information as it is foreseen to be used by specialists who develop lexicons. It includes the following fields, which were already used in one previous work [5]:

1. **Entry** (Entrada): the lexical unit;
2. **Source Language** (etimologia): the language from which the loanword comes from (etymology);
3. **Vernacular form** (étimo): the original lexical unit in the language from which the loanword came from (etymology);
4. **Grammatical Category** (Cat. gramatical): indication of the word class(es);

5. **Variant** (Variante): different spelling variants of the loanword;
6. **Definition** (Definição): Brief and clear explanation of its meaning;
7. **Phraseology** (Fraseologia): allows us to see how the lexical unit occurs in the cotexts;
8. **Image** (Imagem): illustration or figure that illustrates the lexical unit in question.
9. **Remark** (Nota): brief commentary to give an explanation

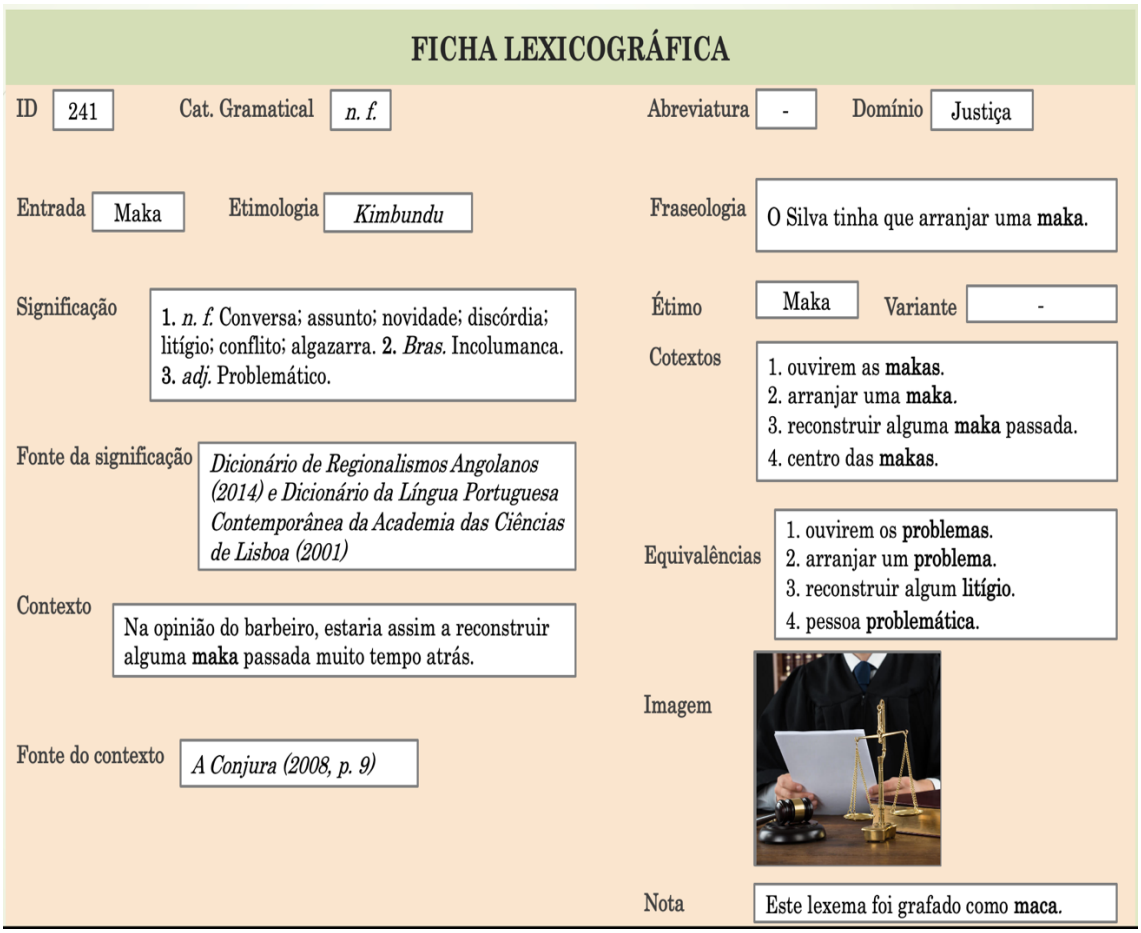


Fig. 8. Lexicographical data sheet for "*Maka*" entry.

5. Conclusions and Future Work

Our aim was to conduct a comparative study of lexical forms to extract lexical loanwords from Angolan languages that are not part of the lexicon of EP with the help of computational methods. We have explained the constitution and characterization of the corpora, which included Angolan found in *A Conjura* (2008), *Jornal de Angola* (2019-2020) and *Telejornal* (2020). We have described the data preprocessing steps done with the help of some Text Mining and Natural Language Processing techniques. We have followed an approach based on the use of the *udpipe* package, which can separate text into tokens in an automatic way, introduce their syntactic categories, and generate the dependencies between them by resorting to dependency parsing.

We have carried out the extraction of loanwords with the focus on loanwords from four lexical classes - verbs, nouns, adjectives, and proper nouns, which originated from several languages of Angola. We have described the criteria used for determining the neological character of lexical units.

With the aid of computational methods, we extracted, by lexical class, 1,784 lexical loanwords from Angolan languages. These are divided into 66 verbs, 244 nouns, 52 adjectives, and 1,422 proper nouns. We verified that a significant subset of these loanwords (923 elements) are not present in existing dictionaries. Furthermore, our contribution here is to present a methodology based on computational methods that allow for processing of other new texts of AP to identify new loanwords. As for the loanwords identified, we made their etymology explicit and clarified their meaning. This was facilitated by the automatic extraction of cotexts with the aid of computational methods.

We structured the loanwords into groups and examine the processes of lexical loan formation. We verified that this study can help the preparation of the common orthographic vocabulary of the International Portuguese Language Institute. Following this analysis, we present the prototype of a dictionary of Angolan regionalisms by means of lexicographical sheets. This work can be continued and extended. In the future, we intend to continue this study by extending the corpus of Angolan Portuguese and, this way, identify other lexical loanwords.

We advocate that lexical forms are the reflection of historical-cultural realities and, as an invaluable heritage, they deserve, from this point of view, our reflection and protection. One line of research could try to determine, or at least suggest, the etymology of loanwords for some languages in an automatic way. It is also possible to study differences in meaning of some words that appear both in Angolan and European Portuguese. This could be done by considering the contexts (represented in the form of graphs or embeddings) in which these words appear and identifying, in particular, significant differences.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCG - Fundação Calouste Gulbenkian.

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within the projects UIDB/50014/2 and UIDB/00022/2020.

References

[1] T. M. C. J. d. Costa, Umbundismo no Português de Angola: proposta de um dicionário de umbundismo, Lisboa: Universidade Nova de Lisboa, 2015.

[2] A. P. G. M. d. Silva, Lexicografia Bilingue de Especialidade: e-dicionário de português-kimbundu no domínio da saúde, Lisboa: Universidade Nova de Lisboa, 2015.

[3] M. Gross, «The Construction of Local Grammars,» chez *Finite-state language processing, Language, Speech, and Communication*, Cambridge, Mass, 1997, p. 329–354.

[4] M. T. Lino et C. Dechamps, «Langue Juridique et Créativité Terminologique: une perspective français-portugais,» *L'innovation Lexicale dans les Langues Romanes,* pp. 83-99, 2016.

[5] T. Muhongo, Empréstimos de Origem Angolana em Voz de Angola Clamando no Deserto, Lisboa: Dissertação de Mestrado apresentada na Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa, 2017.

[6] X. Canosa, X. Varela, P. M. Lema, P. Gamallo, A. J. Taboada et M. Garcia, «Uma Utilidade para o Reconhecimento de Topónimos em Documentos Medievais,» *Linguamática,* pp. 3-15, 2019.

[7] A. Pinto, A. Alves et H. Oliveira, «Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text.,» chez *5th Symposium on Languages, Applications and Technologies (SLATE), vol. 51*, Mariboru, 2016.

[8] P. Gamallo et M. Garcia, «A Resource-Based Method for Named Entity Extraction and Classification,» chez *Progress in Artificial Intelligence*, 15th Portuguese Conference on Artificial Intelligence (EPIA), vol. 7026, Springer, 2011, p. 610–623.

[9] A. Iriguti et V. Feltrim, «Evaluating Features for Rhetorical Structure Classification in Scientific Abstracts,» *Linguamática,* pp. 41-53, 2019.

[10] J. Torruela et J. Llisterri, «Deseño de corpus textuales y orales,» chez *Seminari de Filologia i Informàtica, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona*, Barcelona, 1999.

[11] J.-M. Adam, La Linguistique Textuelle. Introduction à L'anályse Textuelle des Discurs, Paris: Armand Colin, 2005.

[12] J. Sinclair, «Corpus and Text — Basic Principles,» chez *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books, 2005, pp. 1-16.

[13] M. Llamazares, «Lingüística con Corpus,» *Filología,* pp. 329-349, 2008.

[14] R. Dale, «Classical Approaches to Natural Language Processing,» chez *Handbook of Natural Language Processing*, Boca Raton, Taylor & Francis Group, 2010, pp. 3-7.

[15] D. Cielen, A. D. Meysman et M. Ali, Introducing Data Science: Big Data, Machine Learning and More, Using Python Tools, Shelter: Manning Publications, 2016.

[16] S. Caldas, «Lorsque Innovation Linguistique Rime avec Importation Lexicale: Quelques Processus Néologiques d'importation en Portuguais et en Français Contemporain,» *L'innovation Lexical dans les Langues Romenes,* pp. 101-118, 2016.

[17] I. Desmet, «Langues de Spécialité et Foisonnement Néologique en Portugais et en Français: Quelques Réflexions,» *L'innovation Lexical dans les Langues Romenes,* pp. 119-136, 2016.

[18] Ó. Ribas, Dicionário de Regionalismos Angolanos, Lisboa: Mercado de Letras, 2014.

[19] M. A. Mota, «Introdução à Morfologia,» chez *Gramática do Português*, vol. III, Lisboa, Fundação Calouste Gulbenkian, 2013a, pp. 2787-2831.

[20] G. Rio-Torto, «Derivação,» chez *Gramática do Português*, vol. III, Lisboa, Fundação Calouste Gulbenkian, 2013, pp. 3029-3152.

[21] A. Rodrigues, «Introdução,» chez *Gramática Derivacional do Português*, Coimbra, Imprensa da Universidade de Coimbra, 2013, pp. 29-116.

[22] R. Beard, «Derivation,» chez *The Handbook of Morphology*, Oxford, Blackwell, 2001, pp. 44-65.

[23] B. Svensén, A Handbook of Lexicography: The Theory and Practice of Dictionary-Making, Cambridge: Cambridge University Press, 2009.

[24] F. Edmundo, «Sobre a Aprendizagem das Línguas Nacionais, em Angola,» Ciberdúvidas da Língua Portuguesa, Lisboa, 2014.

[25] H. Wickham et G. Grolemund, R for Data Science : Import, Tidy, Transform, Visualize, and Model Data, Sebastopol: O'Reilly , 2017.

[26] S. Weiss, N. Indurkhya et T. Zhang, Fundamentals of Predictive Text Mining, London: Springer, 2015.

[27] J. Pustejovsky et B. Boguraev, «Lexical Knowledge Representation and Natural Language Processing,» chez *Natural Language Processing*, Cambridge, The MIT Press, 1994, pp. 193-223.

[28] P. Murrell, R Graphics, Boca Raton: Chapman & Hall/CRC, 2006.

[29] L. Torgo, A Linguagem R: Programação para a Análise de Dados, Lisboa: Escolar Editora, 2009.

[30] J. Pruvost et J. Sablayrolles, Le Néologismes, Paris: Presse Universitaire de France, 2003.

[31] H. Wickham, Advanced R, Boca Raton: Taylor & Francis, 2015 .

[32] D. Geeraerts, Theories of Lexical Semantics, Oxford: Oxford University Press, 2010.

[33] J. Gama, A. Faceli, A. C. Lorena et M. Oliveira, Extração de Conhecimento de Dados: Data Mining, Lisboa: Edições Sílabo, 2017.

[34] A. C. M. Lopes et C. Carapinha, Texto, Coesão e Coerência, Coimbra: Almedina, 2013.

[35] A. Rodrigues, «Noções Basilares sobre a Morfologia e o Léxico,» chez *Gramática Darivacional do Português*, Coimbra, Imprensa da Universidade de Coimbra, 2016, pp. 35-134.

[36] A. Spencer, «Morphology,» chez *The Handbook of Linguistics*, Oxford, Blackwell, 2001, pp. 213-237.

[37] H. Borer, Structuring Sense, Oxford: Oxford University Press, 2005.

[38] A. Villalva et J. P. Silvestre, Introdução ao Estudo do Léxico: Descrição e Análise do Português, Rio de Janeiro: Vozes, 2014.

[39] M. Haspelmath, Understanding Morphology, London: Arnold, 2002.

[40] I. Mel'čuk, «Morphological Processes,» chez *Morphology. An international handbook on inflection and word formation*,

Berlin, De Gruyter, 2000, pp. 523-535.