Cadernos de
ESTUDOS LINGUÍSTICOS – (59.3), Campinas, pp. 631-648 - set./dez. 2017

ANNOTATING A POLYSYNTHETIC LANGUAGE: FROM PORTUGUESE TO KADIWÉU*

CHARLOTTE GALVES[1]
(UNICAMP & CNPq)
FILOMENA SANDALO[2]
(UNICAMP & CNPq)
TICIANA A. DE SENA[3]
(UNICAMP)
LUIZ VERONESI[4]
(UNICAMP)

ABSTRACT: We propose for Kadiwéu, a polysynthetic language of Brazil, an extension of the POS annotation of the *Tycho Brahe Annotated Corpus of Historical Portuguese* (www.tycho.iel.unicamp.br/~tycho/corpus) – henceforth TBC, which consists in tagging both words and morphemes, yielding a two-level annotation. The tagging of words is necessary to generate the syntactic parsing that is missing from the current corpuses of Brazilian native languages. The morphological tagging is also crucial for polysynthetic languages since it allows searching for grammatical properties encoded by the morphemes. This is a pioneer proposal since it is the first time an American Indian language will be part of a Corpus allowing grammatical searches that include morphological and syntactic information.
Keywords: Kadiwéu, polyssynthesis, Tycho Brahe Corpus, morphological annotation.

1. INTRODUCTION

We propose for Kadiwéu, a polysynthetic language of Brazil, an extension of the Part-of-Speech (henceforth POS) annotation of the *Tycho Brahe Annotated Corpus of Historical Portuguese* (www.tycho.iel.unicamp.br/~tycho/corpus) – henceforth TBC. Kadiwéu is a language from the South of

* Este trabalho foi desenvolvido no âmbito dos projetos temáticos FAPESP 2012/06078-9 e 2012/17869-7. E foi também parcialmente financiado pelo CNPq através de bolsas de produtividade em pesquisa.
[1] charlotte.mgc@gmail.com
[2] fsandalo@gmail.com
[3] ticianaa@gmail.com
[4] luiz.tycho@texugo.com.br

South America that belongs to the Waikurúan linguistic family. It is spoken by about 1,500 Indians distributed over an area of 538,000 hectares in the State of Mato Grosso do Sul, Brazil. The Waikurúan language family has two branches: (a) the Waikurúan Branch, which includes Mbayá and its descendent Kadiwéu; and (b) the Southern Branch, which comprises four other languages: Toba, Pilagá, Mocoví, and Apibón, all spoken in Argentina.

For the TBC, in order to treat the rich inflectional morphology of Portuguese, the POS tagging system of the Penn Parsed Corpora of Historical English (Kroch and collaborators 2000, 2004, 2010) was adapted in such a way that tags can be articulated, with a basic one, corresponding to the category of the word (VB, D, N, NPR, ADJ), and one or more secondary tags which encode morphological properties (-D, -UM-F, -F, -G) (cf. Britto et al. 2002). In a language like Kadiwéu, the information conveyed by morphology is too rich to be treated this way. In these languages, except for some rare cases of portmanteau morphemes and suppletion, the correspondence between form and features is one-to-one, and can be encoded in a single tag. There follows an example of Kadiwéu with morpheme by morpheme glossing.

(1)	ijo	Gonel:egiwa	ja	wajipata.	
	i-jo	*Gonel:egi-wa*	*jaG*	*w-awajipa-ta-wa*	
	D	N		T	VB
	Gnr-Ncl	man-Cla		T	Erg-listen-Obl-Apl
	'The/a man has listened to it.'				

From a linguistic point of view, in a language like Kadiwéu, words and morphemes are treated the same way since the list of tags includes both tags associated with words (in the example above: D, N, VB) and tags associated with morphemes (Erg, Gnr, Obl, Apl, etc). The automatic tagging process will consist of two levels: first, as for languages like Portuguese, the tagger will be run at the level of the whole sentences, in order for each word to be assigned a POS tag. At the second level, the process will be run inside of each relevant word, assigning tags to the morphemes.

The system presented above will be the basis for the syntactic annotation of texts, projected from the word-level tags, as for the languages currently annotated, but without loss of the detailed morphological information typical of polysynthetic languages. This will allow for searches combining the three levels of annotation.

The paper is organized as follows. In Section II, we present the annotation system designed for a Romance language, Portuguese, adapted from the system used in the Penn Parsed Corpora of English. In particular, such adaptation has involved the addition of sub-tags referring to morphology. In Section III, we describe the aspects of the morphology and syntax of Kadiwéu that crucially distinguish this language from Indo-European languages, and will lead us to propose a new annotation system that could be used for corpora of other languages of similar typologies. In this section, we also introduce our tagging system, which departs from what was described for Portuguese in two aspects. First, each syntactic head is assigned a POS tag regardless of whether it is an

independent word or a bound morpheme. Second, we include another layer of annotation that annotates each morpheme of the language. Section IV is devoted to the computational implementation of this annotation system. In Section V, we show how to search in this new system.

2. THE TAGGING SYSTEM OF THE TYCHO BRAHE CORPUS

The tagging system of the Tycho Brahe Corpus (cf. http://www.tycho.iel.unicamp.br/corpus/manual/tags.html) was adapted from the POS tag system of the Penn Parsed Corpus of Middle English (henceforth PPCME) (http://www.ling.upenn.edu/hist-corpora/annotation/index.htm). Because of the richer inflectional morphology of Portuguese, the corresponding tagging set had to be considerably increased, reaching 377 tags instead of 86 for English. In order to avoid the computational complexity deriving from this high number, the tags were split in two parts: a base tag and a sub-tag, separated by a hyphen. Accordingly, the tagging process was split in two steps, one concerning the base tag, and the other the sub-tag (cf. Finger 2000). Table 1 below gives examples of the difference between the tags used in the annotation of the PPML corpora and the tags used in the TBC for verbs other than auxiliaries.

Table 1: *PPCME tagging system vs. TBC tagging system* (*verbs other than auxiliaries*)

	PPCME	TBC	
▲		Without enclitic pronouns	With enclitic pronouns
		pronouns	
Infinitive	VB	VB	VB+CL(+CL)
Inflected infinitive	/	VB-F	VB-F+CL(+CL)
Imperative	VBI	VB-I	VB-I+CL(+CL)
Present	VBP	VB-P	VB-P+CL(+CL)
Present subjunctive	“	VB-SP	VB-SP+CL(+CL)
Past	VBD	VB-D	VB-D+CL(+CL)
Past subjunctive	“	VB-SD	VB-SD+CL(+CL)
Future	/	VB-R	VB-F!CL(+CL)
Future Subjunctive	/	VB-SR	VB-SR+CL(+CL)
Gerund	VAG	VB-G	VB-G+CL(+CL)
Perfect participle	VBN	VB-PP	/
Past participle	VAN	VB-AN	/
-RA verbal forms	/	VB-RA	VB-RA+CL(+CL)

In table 1, we see that the total number of tags for verbs is in Portuguese is 5 times the number of the tags for English (35 vs. 7), since, beyond the fact that Portuguese has a richer tense marking, up to two enclitic clitics can be affixed to the inflected verb. In this case, the TBC annotation system adopted the PPCME use of the + symbol that codifies the contraction of two different words, which have to be split at the level of the syntactic annotation.[5] In 2 we give an example of an excerpt of text annotated along those lines, in which the application of the system can be observed with other categories, like D (Determiners), Q (Quantifiers) and PRO$ (Possessives) can be associated with a gender (-F) and/or number (-P) mark. We also see the numerous cases of contraction of a preposition (P), with a determiner (P+D).

(2) Afirmo/**VB-P** à/**P+D-F** Vossa/**PRO$-F** Excelência/NPR que/C foi/SR-D o/D meu/PRO$ sentimento/N muito/Q maior/ADJ-R-G do/**P+D** que/WPRO sei/**VB-P** declarar/VB ,/, assim/ADV pelo/**P+D** bom/ADJ sucesso/N desta/**P+D-F** minha/**PRO$-F** missão/N depender/VB da/**P+D-F** presença/N e/CONJ autoridade/N de/P Vossa/**PRO$-F** Excelência/NPR ,/, como/CONJS pelo/**P+D** grande/ADJ-G desejo/N que/WPRO eu/PRO trazia/**VB-D** de/P me/CL ver/VB aos/**P+D-P** pés/N-P de/P Vossa/**PRO$-F** Excelência/NPR ,/,reconhecendo-me/**VB-G+CL** Vossa/**PRO$-F** Excelência/NPR pelo/**P+D** seu/PRO$ mais/ADV-R afeiçoado/**VB-AN** e/CONJ mais/ADV-R obrigado/**VB-AN** criado/N ,/, e/CONJ logrando/**VB-G** eu/PRO de/P mais/ADV-R perto/ADV a/**D-F** mercê/N que/WPRO Vossa/**PRO$-F** Excelência/NPR em/P toda/**Q-F** parte/N há/HV-P sido/SR-PP servido/ADJ fazer-me/**VB+CL** ./. V-002,0.3/ID ./PONFP

This system already represents a step forward the annotation of a richer morphology than English. As concerns a polysynthetic language like Kadiwéu, however, it is not adequate to provide us with a computationally exploitable system of annotation able to encode the huge richness of the morphological component. As we shall see in the next section, verbs in this language can bear up to 14 morphemes, realized as both prefixes and suffixes. In this case, the number of POS tags would increase in such a way that such a system could barely be computationally implemented. As an example, take the case of nouns and verbs. In Portuguese, the longest possible tag is composed of 2 elements for nouns (N-P), and of 4 elements for verbs (VB-P+CL+CL). In Kadiwéu, we would have to build POS tags with up to 7 elements for nouns and to 14 elements for verbs in as many combinations as possible (cf. Section IV). Moreover, the structure of the tags would be more complex since some of the sub-tags would precede the basic tag and some other tags would follow it.

[5] The ! symbol is an innovation of the TBC that codifies the mesoclitic position of clitic pronouns when the verb bears future tense.

In the next section, we present an overview of the morpho-syntax of Kadiwéu, which we will guide our tagging proposal.

3. AN OVERVIEW OF THE MORPHOLOGY AND SYNTAX OF KADIWÉU AND OUR TAGGING PROPOSAL

Morphologically, the noun, verb, and D-system (Ds, quantifiers, numerals, interrogative pronouns) are highly inflected.

3.1.3. Verbs

The verb agrees with the subject and object in person and number (sg/pl). Person and number are not grammaticalized together in Kadiwéu, however (except for 1pl direct and indirect objects). So, we need independent tags for person and number.

Kadiwéu person is marked by a set of ergative agreement prefixes for transitive and inergative subjects and a set of absolutive agreement prefixes for direct objects and intransitive subjects. Thus, the verb shows agreement with internal and external arguments, but the subject and object agreement markers are in complementary distribution according to a hierarchy 1PLOBJ>2>1>3. We indicate whether the agreement is ergative or absolutive in our tagging since just one of them will be marked as a prefix. There is also a prefix for an impersonal subject.

Indirect objects as well as any oblique argument are marked by agreement as well. They are suffixes.

Kadiwéu has an inverse voice system and the verb is marked by an inverse morpheme whenever there is fronting of an internal argument (that occurs with any second or first-person internal argument (due to an inverse voice), unaccusative and reflexive verbs, among other fronting structures).

The verb is also marked by applicative, directional and motion markers, and aspect morphemes. The verb roots does not convey motion and directionality in these languages. And the language lacks adpositions as discussed later. Applicative verbal suffixes appear whenever a language as Portuguese or English would have prepositions.

Finally, there are valence changes morpheme that include antipassive morphology and other increasing or decreasing valence morphemes.

The morphological tags associated with verbs are presented in the table below, followed by examples. Note that the root is marked by v; in Kadiwéu there is no uninflected verbal root.

Table 2: *The verb structure.*

POS TAGS	Morpheme Tags		Examples	
VB	Plu	plural	o-y-a:lGe Plu-Erg-v	'They kidnap him.
	Imp	impersonal	eti-Ga-d:-d:egi Imp-Abs-Inv-v	'Someone brought you.'
	Erg	ergative agreement	j-awi: Erg-v	'I hunt it.'
	Abs	absolutive agreement	i-d:-abi-d Abs-Inv-v-Asp	'I'm standing up.'
	Inv	inverse voice	Go-d:-ili: Abs-Inv-v	'We grow.'
	Ant	antipassive	n-ema-ta Ant-v-Obl	'She/he loves him/her in distance.'
	Hit	hither	n-ad:e:gi Hit-v	'He brings it.'
	v	verbal root		
	Val	valence change morpheme	j-otaGan-Gen:-aGa Erg-v-Val-Plu	'We talk to him.'
	Asp	aspect	o-y-aqage-di Plu-Erg-v-Asp	'They cut it.'
	Obl	oblique argument agreement	me-ta v-Obl	'He says to him.'
	Dir	directional morpheme	ji-l:o-ko-tigi Erg-v-Val-Dir	'I look up at something.'
	Mot	motion	ji-n-otiqo-tijo Erg-Hit-v-Mot	'I come wistling.'
	Apl	aplicative	j-ao-tGa-domi Erg-v-Obl-Apl	'I make it for you.'

3.2.3. Nouns and D-words.

The noun is obligatorily inflected by possession. Alienable nouns must additionally be marked by an antipassive morpheme. Only borrowed nouns, unpossessed nouns like names that refer to nature, and proper names, are bare.

The noun can be optionally inflected by diminutive, number, and gender. Number and gender, however, are obligatory in the determiner system.

Kadiwéu, like other Waikuruan languages, is a deictic classifier language (Aikhenwald 2000). Sandalo (2015) has interpreted deictic classifier languages as numeral classifier languages in which numeral classifiers are an obligatory ingredient of all determiner-like elements, such as quantifiers, numerals, and wh-words for arguments. These D-words have a complex morphology, as it can be noted in the Table 3 below. Bare nouns are normally interpreted as number neutral: in the case of count nouns, a bare singular NP is interpreted as a group (of 1 or more representatives of the kind), while a bare singular mass noun is interpreted as an unspecified amount of a substance. Even bare plurals of count nouns are interpreted as denoting several groups rather than individuals. Once a classifier is present, count nouns are seen as atoms in the singular, and as more than one individual in the plural, and masses are necessarily interpreted as packaged/ coming in (a number of) containers. Finally, the determiner system can also mark anaphoricity (i.e. whether a noun has been mentioned before in a text or is part of a common knowledge).

Additionally, there are classifier morphemes that denote shape and other properties of a noun and derivational morphemes that generate deverbal nouns.

The morphological tags corresponding to nouns, demonstratives, numerals, and quantifiers are below:

Table 3: *The Noun and D-words structure.*

POS TAGS	Morpheme Tags		Examples	
N	Gen	genitive agree-ment	l-okaGe-te-di Gen-n-Cla-Plu	‘his friends’
	Ant	antipassive	n-gato-je Ant-n-Cla	‘a bullet’
	n	root	dom: o:jya n	‘car’
	Cla	classifier	apaqa-co-di n-Cla-Plu	‘rheas’
	Der	derivation	n-dele-Gikajo Ant-v-Der	‘warrior’
	Dim	diminutive	l-atope-nig: i Gen-n-Dim	‘his gun’
	Plu	number	Gonel:egi-wa-tedi n-Cla-Plu	‘groups of man’

D				
	Anf	anaphoric	nG-i-jo nG-Gnr-Ncl	This/the/An one_ mentioned before
	Gnr	gender	i-di Gen-Ncl	This/the/An one
	Ncl	numeral classifier	i-di Gen-Ncl	This/the/An one
	Plu	number	i-di-wa Gnr-Ncl-Plu	These/the ones
NUM	Num	numeral	i-ni-wa-ta:le Gnr-Ncl-Plu-Num	two
Q				
	Qnt	quantifier	oni-ni-te-k-beke Num-Gnr-Ncl-Obl-Apl-Qnt	each
WPRO	Int	interrogative pro-noun	am-i:-na Int-Gnr-Ncl	'who'
WADV	Whs	Wh-support	ig-ame Whs-Int	'why'
PRO	Pro	pronoun	aqa:m:-i Pro-Plu	'you'

3.3.3. POS tags and phrasal labels

In spite of the fact that the Kadiwéu POS system is an extention of the POS system used for Portuguese in the *Tycho Brahe Annotated Corpus of Historical Portuguese*, the grammatical differences between the two languages oblige us to introduce a few changes in the tagging system for Kadiwéu.

First, in Kadiwéu, aspect and mood are not marked by bound morphemes on the verb but by independent functional words. Kadiwéu has seven aspect markers --- completive/incompletive/durative, telic/atelic, repetitive, and intensive --- and two mood markers, conditional and desiderative. This is why we have decided to have T and MOD as independent POS tags.

Second, Kadiwéu has a quite recursive genitive grammar, in which each noun is marked as possessive. We assign a special tag to these nouns, N$, which is not used in the Portuguese system but already exists in the English system.

(3) Ganeb:i wa:ka libol:e libinyenig:i

N$ N$ N$ N$

'Your beautiful cow's meat' (Lit.: your cow its meat its beauty')

Complement clauses are introduced by the complementizer *me*. Control structures --- structures in which either the subject of the main clause is also the (semantic) subject of the subordinate clause or the object of the main clause is also the (semantic) subject of the subordinate clause --- have the same structure. That is, the main and subordinate clauses must be separated by the complementizer *me*.

Adverbial clauses are introduced by *nige*, *naGa*, and *noaGa.* The complementizer *nige* can be glossed as 'C (fut)', *naGa* as C (non-fut)', and *noaGa* as 'where'.Tensed complementizers will be tagged CT and untensed complementizer (including the one that introduces relative clauses) will be tagged C. Relative clause are introduced by the complementizer *ane, which* is also tagged C. Location, purpose, and manner clauses also involve relativization. They are introduced by a relative pronoun followed by another C node: *i* 'introduces place clauses', *le:Godi* 'introduces explicative clauses', and *oda:Ge:* 'manner clauses'.

Kadiwéu lacks prepositions entirely. The structures that contain prepositions in languages like English have applicatives (therefore we have included a VBAPL tag) or are in a serial verb construction in Kadiwéu.

As Portuguese, Kadiwéu has a rule of contraction between two adjacent words, the first of which corresponds to a functional category. Since each word has a separate syntactic function, they have to be split at the syntactic level. As for Portuguese (cf. Table 1), we use in this case a complex tag in which the + symbol indicates that the corresponding morphological unit is composed of two independent heads. The complete list of POS tags is given in Table 4 below. The tags that are not part of the tagging system for Portuguese are marked with an asterisk.

Table 4: *POS tags for Kadiwéu*

POS Tags	Examples
ADJ	icoa Goneleegiwa ejiwajegi D N ADJ 'a kadiweu man.'
ADV	igotibeki ditigedi VB ADV 'he goes far.'
*AUX	ejigo jaaGadi nigatoje gaxiana AUX VB N$ N 'I am going to finish the bullets from Paraguayans.'
C	jatita napalite me jaqagidi VBAPL N$ C VB 'I use a machete to cut it.'

C+C	
C+D	miniwa me GodapoaGeneGegi one agelexaGa C+D C N$ VB N ‘The warrior was said to be paunchy.’
C+NEG	Pedro meta Paulo medaGa dinojeteta dom:o:jya NPR VBAPL NPR C+NEG VBAPL N ‘Peter told Paul not to buy a car.’
C+N	pida diGawini anoko ejewajegi moko otimadaGe CONJ VB C+N ADJ C+N ADJ ‘But look at this kadiweu people, who are stubborn people.’
C+AUX	jonaGa node icoa lokaGetedi migotibece oideleGe gaxiana T VB D N$ C+AUX VB N ‘He invited his friends to fight against Paraguayans.’
C+VB	
C+VBAPL	me:tGawa migo aqi:di VBAPL C+VB N ‘He said to you that he goes to the river.’
	icomGated:ijo itwata:l:e noqododi monipaditegi Ecabigo. VBAPL NUM N C+VBAPL NPR ‘They have been waiting for Ecabigo for two days.’
	icomGated:ijo itwata:l:e noqododi monipaditegi Ecabigo. VBAPL NUM N C+VBAPL NPR ‘They have been waiting for Ecabigo for two days.’
*CT	niGijo niganigawanigi naGa diote D N$ CT VB ‘When the boy slept.’
CONJ	pida ayema ica me diote CONJ NEG+VB D C VB ‘But he did not want to sleep.’

D	ica begi D N ‘a hole
INTJ	oda one ewo ica layageGegi: toli toli toli CONJ VB VB D N$ INTJ INTJ INTJ ‘His noise was said to be: toli toli toli.’
* MOD	domaGa yema João MOD VB NPR ‘She is about to love John’
*MOD+VB	dGaid:inicitike bitGa id:oy MOD+VB N VB ‘If I swing myself, I feel fear.’

N	ica apakanigo D N ‘a rhea’
*N$	dinatopeteloko lakilo. VBAPL N$ ‘He shot himself in the head.’
NEG	Pedro i:Ge iwal:o me daGa yad:e:gi naqakodiwaGa. NPR VB N C NEG VB N$ ‘Peter ordered the woman not to take away the rice.’
NEG+ADV	leeGodi aGale oyatopetelogo CONJ NEG-ADV VBAPL ‘Because they no longer shot.’
NEG+AUX	aGejigo jawi: NEG+AUX VB ‘I am not going to hunt.’
NEG+VB	Pedro ayema: me dawi:. NPR NEG+VB C VB ‘Peter does not want to hunt.’
NEG+VBAPL	Maria adatiqata nod:a:jo oqoqo:di me yel:wadi oqoqo:di. NPR NEG+VBAPL N$ N C VB N ‘Mary killed the chicken not with her knife.’
NPR	Joao aja Maria diniqe: NPR CONJ NPR VB ‘John and Mary introduce each other.’
NUM	itiwata:le laketedi NUM N$ ‘two snakes.’
PRO	aqa:m:i Gad:ma:n:i PRO VB ‘He loves you.’
Q	aGaca gatooje Q N ‘no bullet’
T	oda jonaGa ibagi CONJ T VB ‘Then he went to sleep.’
T+N	jiGijeGagi T+N ‘It has already become a wild animal.’
T+AUX	joneGigo igo T+AUX VB ‘He was going.’

T+VB	jonoGonadi ica apakanigo T+VB D N 'He saw the rhea.'
T+VBAPL	AneotedoGoji joneGeniditeta icoa liwikatedi NPR T+VBAPL D N$ 'God called his dogs.'
VB	inodenGa VB 'We invite him.'
*VBAPL	jajigotGawa lib:ol:e VBAPL N$ 'I give meat to you.'
WADV	igamei Pedro me yoe ika di:m:igi? WADV NPR C VB D N 'Where did Peter make a house?'
WPRO	ami:jo ica ane din:ojete:ta id:a? WPRO D C VBAPL D 'Who is buying it?'

Note that if we sum the two set of morphological tags, proposed above in Tables 2 and 3, with respectively 12 and 15 elements, plus the POS tags listed in Table 4, which totalizes 37 elements, we obtain the total of 64 tags. This is a rather economical system, if we compare it to English (around 80) and Portuguese (more than 350). This again suggests that our proposal of splitting the POS and the morphological level annotation is on the right track. Moreover, only 6 basic POS tags (plus 2 complex tags) are not in the list currently used for English or Portuguese (see the asterisks in Table 4). This is a very nice and encouraging result with respect to the feasibility of the adaptation of already existing annotation systems to entirely new languages.

As for the syntactic annotation of Kadiwéu, things are even simpler, since we can simply adopt, as a point of departure, the whole list of categories used to annotate English and Portuguese. (cf. respectively http://www.ling.upenn.edu/hist-corpora/annotation/index.htm and http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/syn-frm.html). We leave as an empirical question the rigorous definition of the syntactic labels necessary to annotate the language. As a first approximation it is possible to claim that some of the existing ones, as expected, will not be relevant due to some specific grammatical properties of the language. First, as already mentioned, P does not exist as a category and therefore, no kind of PP is projected. Second, not all the types of IPs (Inflectional phrases - a label referring to either clauses with no complementizer, or to the part of the clauses governed by complementizers) used in the English and Portuguese annotation system are necessary to describe Kadiwéu syntax. For instance, finiteness is not expressed in clauses. Therefore, the categories denotating infinitival, gerundive and participial clauses in English and Portuguese, respectively IP-INF, IP-GER and IP-PPL, will not be required, for Kadiwéu. The main distinction will be therefore between matrix clauses (IP-MAT) and subordinate clauses immediately dominated by CP (IP-SUB).

We leave for further research the question of whether the notion of small clauses (IP-SMC) is relevant for the annotation of this language.

As for CPs, Kadiwéu has, like English and Portuguese, complement clauses (CP-THT), adverbial clauses (CP-ADV), interrogative clauses (CP-QUE), relative clauses (CP-REL, CP-FRL), comparative clauses (CP-CMP), and cleft sentences (CP-CFL) .

Finally, Kadiwéu has, as expected, all the functions expressed by NPs in other languages: subjects (NP-SBJ), direct objects and nominal predicates (NP-ACC), genitive noun phrases (NP-GEN), appositive noun phrases (NP-PRN), and adverbial noun phrases (NP-ADV). We shall only add the label NP-APL for the equivalent of PPs in Portuguese and English. It is a matter of further research whether dative objects are to be labeled independently, using the NP-DAT tag.

4. SYTEM IMPLEMENTATION

In the current system of the Tycho Brahe Corpus, the tagging step and the parsing step are completely dissociated. The tagger[6] is incorporated into the xml editor e-dictor, while the parser[7] runs on the txt output of the tagger, creating a file which is completely independent of the rest of the system. We are now implementing a new system (the Tycho Brahe Framework) in which all the steps of the processing of the texts are integrated with each other.

In this section we detail the computational implementation for the tagging and parsing processes integrated into this framework, a web-based system composed of a complete system of edition and annotation called eDictor[8] and search mechanisms for morphological and syntactic purposes cf. www.tycho.iel.unicamp.br/tbf/.

Basically, the system operates in a multi-corpora environment that allows multiple corpora to be maintained all together with multiple available services stored in the cloud. eDictor is the main interface for user input. The information must be presented to the software in a way that allows the user to insert textual, morphological and syntactic information.

The first step is for users to inform which are the morpheme or/and POS tags that they will use in their annotation schema.

[6] We currently use Fabio Kepler's tagger.

[7] We currently use Dan Bickel's parser.

[8] Paixão de Sousa, M. C. et al. e-Dictor: novas perspectivas na codificação e edição de corpora de textos históricos. 2010.

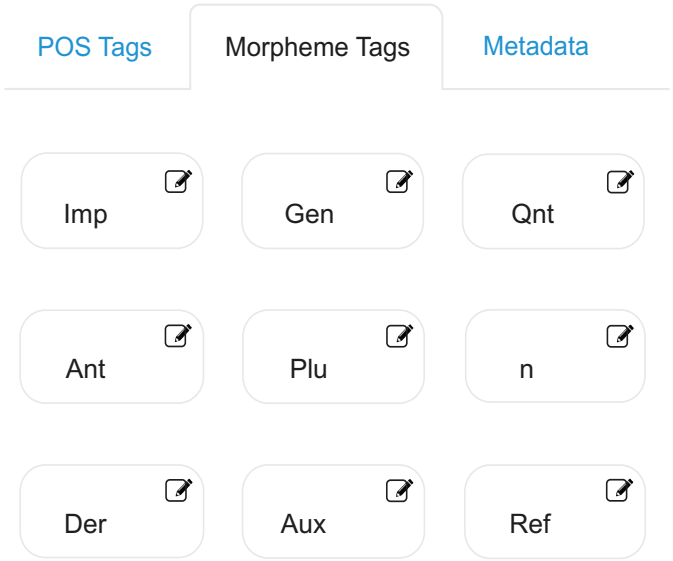


Figure 1: *Selection of the morpheme tags for Kadiwéu*

Those must be predefined for the software to be able to run tagging and parsing. With that in mind, the eDictor interface should be able to let users choose which tags are appropriate for each word or morpheme from this predefined set.

In the case of a language like Kadiwéu, where morphemes are tagged independently from words, the mechanism is the following. When users click on the word, a box is opened for the user to inform the morphemes that compose that word. This is done by simply inserting a white space among each morpheme.

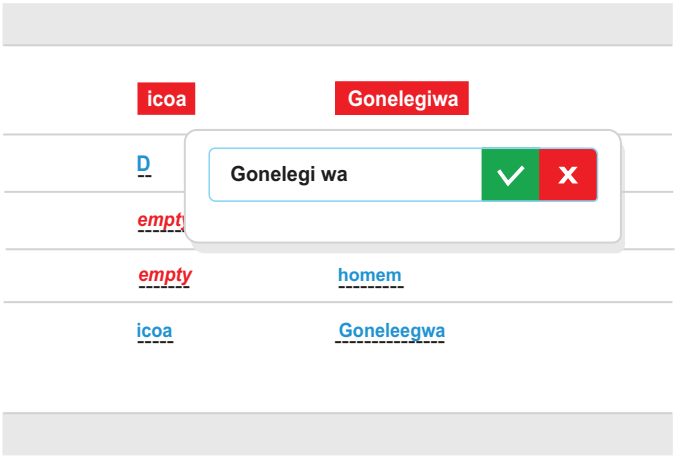


Figure 2: *Morpheme tokenization.*

The system, then, catches this string and breaks it by applying the tokenization process, creating a computational object for each morpheme in the document structure.

Visually, words (tokens) and morphemes (split tokens) are presented to the user with different colors and, below each, there is a field where the user can choose the appropriate tag for that token or split token. The system automatically feeds the list with the correspondent tags, i.e., if it is a token, only the POS tags are available, otherwise, if it is a split token, the morpheme tags are displayed. The interface also has a button to toggle between split tokens and tokens, so that it is possible to set a POS tag to a token and different morpheme tags for each split tokens related to that token. This is what we call a two-level tagging approach.

Figure 3: *Token edition.*

Internally, the system is organized in a system of parameters. The parameter, for the system, is a typified entity, which means that this entity is designed as a single computational object for several purposes. In this case, one type is the POS tags and the other is the morpheme tags.

The way parameters are used depends absolutely on the logic set for the system. The POS tags and morpheme tags are used within the system for three major reasons :

1 - To display the list of tags at the interface for the user (as explained above),

2 - To distinguish different tokens for the corpus dictionary,

3 - To feed the tagger and parser trained models.

The Tycho Brahe Framework has also a built-in dictionary model for each individual corpus. This means that each word or morpheme (token or split token) is recorded into the database to be used for two main purposes: to fill token information automatically at the interface and to realize automatic linearization. POS tags are used within a pair-based approach to keep token singularity, i.e., the system conceives each word as a pair of its value (the word per se) and its tag.

All this information feeds the tagger and the parser trained models. The system uses this stored information to keep track of the most probable combination of words given by the user as its input. As a probabilistic model, the tagger 'guesses' which are the most probable tags for each given word. One important thing to notice here is that the tagger needs some primary input information to be able to analyze the words and return the associated tags. This is why there is an interface for receiving this information. The tagger and parser accuracy directly depends on the amount of given information.

This way of working with two independent tagging levels allows the system to behave in two different ways with tags. First using POS tags to feed the tagger as explained above and second to realize linearization (automatically split tokens into smaller units) associating the correspondent morpheme tags with the words.

5. SEARCHING

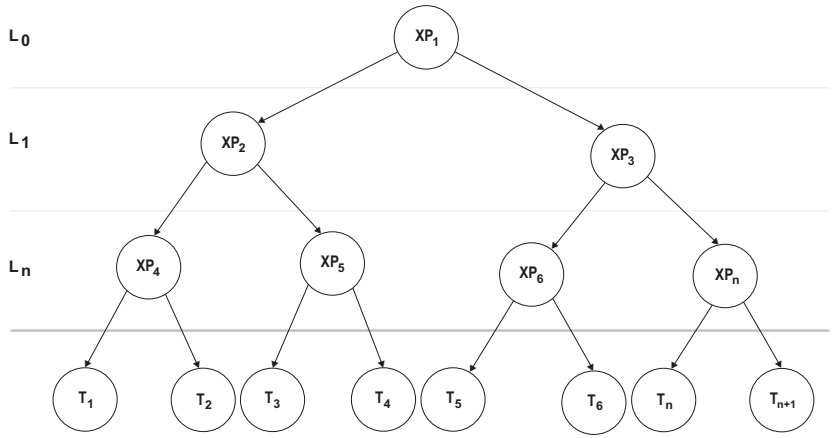
The Tycho Brahe Framework will provide two different approaches for searching over documents and corpora. The first one is a word/POS tag linear mechanism that allows the user to search for any word w.r.t either its specific value or its POS tag, i.e., the user may search for a specific word at a specific position in the sentence, for example, all sentences that starts with the word 'Hello' and realize combinations with more words, for instance all sentences that starts with the word 'Hello' followed by the word 'world' or all sentences that starts with the word 'Hello' followed by any word tagged with the POS tag 'N' for noun.

Through the interface, the user creates his queries by associating boxes that allows him to select a word ('Hello') or a POS tag (the list of available tags is displayed). He may also search for specific morphemes or morpheme tags combining with the words, for example, all sentences that starts with the word 'Hello' and whose next word is composed by this or that morpheme. This mechanism is still under development and will be released soon.

Figure 4: *Linear search interface (still in development)*

The second approach consists in a purely syntactic search where the system's interface allows the user to draw a piece of a syntactic tree starting from the syntactic nodes like IP, CP, NP, etc. down to the most embedded words.

This mechanism consists in decomposing the tree into several smaller chunks as exemplified below.



chunks	*level*	*initial token*	*final token*
chunk XP_1	L_0	T_1	T_{n+1}
chunk XP_2	L_1	T_1	T_4
chunk XP_3	L_1	T_5	T_{n+1}
chunk XP_4	L_n	T_1	T_2
chunk XP_5	L_n	T_3	T_4
chunk XP_6	L_n	T_5	T_6
chunk XP_n	L_n	T_n	T_{n+1}

Figure 5: *Hierarchical search interface (still in development)*

After that, all chunks are converted into a query for searching over the structures stored into a NOSQL database to retrieve all sentences that match the criteria given by the user. This mechanism is also under development and will be released soon.

Both mechanisms store all queries made by the user for future research and are also integrated into an analytics module based on document metadata information like geographical and period information, literary genre and others.

Another important feature for those mechanisms is that they allow the users to continue searching over the results of a searching as many times it is possible and also choose which mechanism to use, for example, the user may first realize a linear searching then a syntactic one over the results and vice-versa.

The Tycho Brahe Framework is still under development but it can be visualized at the following address: http://www.tycho.iel.unicamp.br/tbf

REFERENCES

AIKHENWALD, A. (2000). *Classifiers: A Typology of Noun Categorization Devices*. Oxford: Oxford University Press.

BRITTO, H., Finger, M., GALVES, C. (2002). *Computational and linguistic aspects of the construction of theTycho Brahe Parsed Corpus of Historical Portuguese*. Romance Corpus Linguistics - Corpora and Spoken language. Tubingen: Narr.

FINGER, M. (2000). *Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe*. Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR2000).

SANDALO, F. (1995). *A Grammar of Kadiwéu*, Unpublished PhD dissertation. University of Pittsburgh. Sandalo, F. (2015).