A Construction Grammar Approach for Pronominal Clitics in European Portuguese

Tânia Marques and Katrien Beuls

School of Informatics
University of Edinburgh
Edinburgh EH8 9AB
United Kingdom
tmarques@inf.ed.ac.uk

Artificial Intelligence Lab
Vrije Universiteit Brussel
Pleinlaan 2 B-1050 Brussels
Belgium
katrien@ai.vub.ac.be

Abstract. Cliticization in European Portuguese (EP) is unique amongst other Romance languages. While preverbal and postverbal placement of clitics is common, it is defined by the finiteness of the verb. In EP, however, the clitic placement does not depend on the verb, but on the context surrounding it. In this paper, we present an operational construction grammar model for parsing and production of proclitic contexts.

Keywords: Construction Grammar, Pronominal Clitics, European Portuguese, Language Comprehension, Language Production

1 Introduction

In European Portuguese (EP), as in other Romance languages, pronominal clitics can appear before the verb (proclisis) or after the verb (enclisis). However, in other languages, the type of cliticization is usually defined by the finiteness of the verb [1]. In EP, cliticization is more complex, because the same verb form can appear attached by the same pronominal clitic preverbally or postverbally, depending on the context. This is illustrated by sentence (1), where proclisis appears because of the conjunction 'porque' (because) before the verb in the second clause, while no trigger is present in the first clause, leading to an enclitic.

(1) Eu dei-te este livro, porque tu o querias
I gave 2SG.DAT this book, because you 3SG.MASC.ACC wanted
'I gave you this book, because you wanted it'

A solid theory for proclisis is essential for creating computational models that are able to produce sentences (i.e. generated from meaning) in EP. Existing parsers (e.g. LXParser [2], MSTParser [3], Palavras Parser [4]) have no problem identifying clitics which have more or less unique forms and are usually adjacent to the verb. But a system for producing sentences is dealing with far fewer hints about the placement of the clitics and is therefore likely to over-generate.

Luís and Otoguro [5] have identified proclitic contexts that could inform computational systems for language production. In their work, the position of the clitic and the triggers are defined through functional precedence relations, using

Lexical-Functional Grammar (LFG) [6]. However, the argument structures and the constructions proposed have not been implemented in concrete operationalized patterns, at least as far as we know. Inspired by Luís and Otoguro's theory [5], we present an initial computational model, where the full cycle of parsing and producing sentences is used as an evaluation mechanism. Fluid Construction Grammar (FCG) [7, 8] is used instead of LFG, because of its partial mapping between syntactic and semantical arguments. This increases its flexibility when dealing with unknown arguments, in contrast with LFG which assumes that each lexical entry dictates all necessary linking information [7].

Our approach relates to semantic parsers, because it attempts to retrieve a meaning to enable production of the correct sentence back. Our initial grammar is based on syntactical phrasal templates obtained from the proclitic contexts found in the literature. Nevertheless, it could be further extended by introducing new constructions and lexicons. As in the work by Wang et al. [9], our model is fully functionality driven and could be easily adapted to a crowdsourcing model for automatic expansion. As stated by the authors, this has the advantages of not requiring annotation from error prone sources, and of enabling the system to learn a lot of the language variability with a small number of utterances.

2 Modeling a Fragment of Portuguese Grammar

Our grammar is implemented in Fluid Construction Grammar (FCG) [7, 8]: a formalism that represents linguistic information by means of feature structures. Both for basic units (e.g. noun), or more complex structures, created by matching basic units. It works for strings (when parsing) and meaning networks (when producing). For more information, please consult Steels [8].

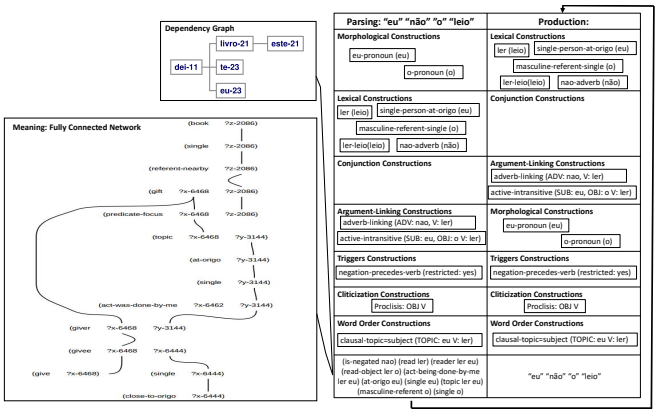

Fig. 1. Diagram with the order of application of the constructions in parsing (left), and in production (right). The meaning obtained in parsing is used as input in production.

In our grammar, for simplification, we ignore all morphological changes that occur in enclisis and mesoclisis, and we do not accept auxiliary verbs.

Figure 1 gives an overview of our grammar[1]. There are six types of constructions which apply in different orders in parsing and production. In parsing, morphological constructions are the first to apply by matching to the strings in the input, then semantic and syntactic features are added by the lexical constructions. If one of the words is a conjunction, then the sentence is divided into two clauses by the conjunction constructions. Argument linking constructions identify the semantic connections between event participants and build grammatical dependencies. Trigger constructions find triggers to restrict the position of the clitic by cliticization constructions. Finally, word-order constructions take care of the remaining orderings between words by looking at the topic and/or focus of the clause. The output of parsing will be a fully connected meaning network that is used to generate the sentence back in production and a dependency graph with the connections amongst dependents. Production is similar to parsing, but lexical constructions are applied first by matching with the meanings, and the morphology of words is decided only after the argument linking constructions.

3 Proclitic Contexts

Fronted Focus: Clitics appear preverbally if the focus comes before the verb (see (2) and (3), taken from Luís and Otoguro [10], which differ only in the focus). This happens when the first position of the clause (topic) is not the subject. The topic of a sentence is defined by a meaning predicate called 'topic' that is by default linked to the subject argument of the main verb. It is worth noting that the discourse topic is not the topic of the clause. In 'Esse livro, li-o eu', 'Esse livro' is the discourse topic, but not the topic of the second clause, which is empty. Furthermore, the focus 'eu' is not fronted, not leading to proclisis.

(2) Eu dei-te este livro
I gave 2SG.DAT this book
'I gave you this book'

(3) Este livro te dei eu
This book 2SG.DAT gave I
'I gave you this book'

This behaviour is achieved by two constructions: one that says that the clausal focus is not the subject, and adds a (restricted +) feature to the verb, activating proclisis; and the other one identifies the predicate focus (well-known concept in linguistics [11]), which moves the subject to the back of the sentence.

Wh Questions: Wh questions that are not the subject and precede the verb will automatically trigger proclisis, since they are fronted focus. For dealing with Wh questions that are subjects, an additional construction wh-fronted, was introduced that adds (restricted +) feature to the verb when a Wh question is identified before the verb. The wh-in-situ was also introduced, which adds a (restricted -) to the verb, allowing for Wh questions in the end of the sentence.

(4) Quem te deu este livro?
Who 2SG.DAT gave this book
'Who gave you this book?'

(5) Tu deste-o a quem?
You gave 3SG.MASC.ACC to whom
'You gave it to whom?'

Negation and Adverbs: Most adverbs can either precede the verb (7), or come in the end of the sentence (8). This is not true for negation (6), because it

[1] A demo of the grammar can be found in http://fcg-net.org/demos/propor-2016/

always comes before the verb. Furthermore, there is a class of adverbs that do not trigger proclisis even when preceding the verb as in (8). Luís and Otoguro [5] classify the proclisis triggering adverbs as operator-like. In our implementation, the constructions look for an adverb that precedes the verb and it is operator-like or negation, which we use to restrict the verb. There is also a construction that accepts adverbs after the verb, if these adverbs can come postverbially.

(6) Ele não te deu este livro
He not 2SG.DAT gave this book
'He did not give you this book'

(7) Eu raramente o leio
I rarely 3SG.MASC.ACC read
'I rarely read it'

(8) Eu leio-o raramente
I read 3SG.MASC.ACC rarely
'I read it rarely'

(9) Eu ontem vi-te
I saw 2SG.DAT
'I saw you yesterday'

Quantified Subjects: Proclisis can also happen when specific quantified subjects precede the verb. This is not true for all quantifiers, though. The quantifier 'algumas' (some) is not a trigger (9), while 'poucas' (few) is (10). Crysmann [12] classifies the proclisis triggering quantifiers as 'downward entailing quantifiers', because they seem to have downward monotonicity. Following this idea, we added a feature *(downward +)* that is only present in lexical constructions for those quantifiers. Thus, the trigger construction looks for a 'downward entailing quantifier' that is quantifying an element related to the verb and precedes it.

(9) Algumas pessoas lêem-no
Some people read 3SG.MASC.ACC
'Some people read it'

(10) Poucas pessoas o lêem
Few people 3SG.MASC.ACC read
'Few people read it'

Subordinate Conjunctions
Complementarisers, relative clauses and subordinate conjunctions will also trigger proclisis. Here, we will focus on subordinate conjunctions, since they are all similar contexts. When a subordinating clause is connected to a subordinated clause, there are two different verbs in the sentence. So, our method for individualizing the clauses is by identifying the conjunction and the verb in each of them, and introducing the feature (restricted +) to the verb of the subclause.

(11) Eu dei-te este livro, porque tu o querias
I gave 2SG.DAT this book, because you 3SG.MASC.ACC wanted
'I gave you this book, because you wanted it'

4 Analysing the Model

All our examples and further sentences were tested to assure that they can be parsed and their meaning can produce exactly the same sentence, i.e. does not overgenerate. Yet, a more through evaluation of the system is needed. Usually, this is done with a well-known annotated corpus. In our case, such standard evaluation is unfeasible, because there are syntactic patterns that are unsupported, and because the required information differs from standard corpus. A solution would be to create our own corpus, but it would be inherently biased. Instead, we opted to describe the sentences that can be produced. Table 1 contains 18 templates. Each template was evaluated in terms of how many variations of a sentence can be generated (expected complexity column). The grammar requires grammatical agreement of gender and number. We consider 2 types of

article (definite/indefinite with each 4 forms), 3 proximities in demonstratives, 8 nominative personal pronouns, 11 clitics (reflexives, accusatives and datives), 6 of which are datives. With the different orders that the elements can take, we expect our system to accept and generate $n_{sentences}$ sentences which could be obtained by automatically introducing morphological and lexical constructions.

Overgeneration of our model is limited by the grammatical agreement, but more importantly, by the meaning networks that directly steer the production process. The system mostly overgenerates in the word order. We could be more strict, but it would not capture the order flexibility in EP. Because of the lack of a grammaticality criterion, ungrammatical sentences are sometimes processed coherently, meaning that the parsing process resulted in a fully connected meaning network and the produced utterance is actual a correction for the original input sentence. This robustness check is an advantage in error-prone discourse.

Structure	Examples	Expected Complexity
Noun Phrase		$np = \sum_{i=1}^{4} np_i$
NOUN	Livros	$np_1 = n_{noum}$
Art NOUN	O livro	$np_2 = 2 \times n_{nouns}$
DEM NOUN	Este livro	$np_3 = 3 \times n_{nouns}$
NPron	Eu	$np_4 = 8$
Clauses		$n_{clauses} = \sum_{i=1}^{18} cl_i$
1:NP1 V NP	Eu leio livros	$cl_1 = 2 \times np \times n_{cj\_vb \wedge pr=np1\_pr}$
2:NP1 V CL	Eu leio-o	$cl_2 = np \times n_{cj\_vb \wedge pr=np1\_pr} \times 11$
3:NP1 V CL NP	Eu leio-te este livro	$cl_3 = np \times n_{cj\_vb \wedge pr=np1\_pr} \times 6 \times np$
4:NP CL V NP1	Este livro te dei eu	$cl_4 = np \times n_{cj\_vb \wedge pr=np1\_pr} \times 6 \times np$
5:NP1 ADV CL V	Eu não te vejo	$cl_5 = np \times n_{op\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 11$
6:NP1 ADV CL V NP	Eu não te leio o livro	$cl_6 = 2 \times np \times n_{op\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 6$
7:NP1 ADV V CL	Eu ontem vi-te	$cl_7 = np \times n_{nop\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 11$
8:NP1 ADV V CL NP	Eu ontem li-te este livro	$cl_8 = 2 \times np \times n_{nop\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 6$
9:NP1 V CL ADV	Eu vejo-te raramente	$cl_9 = np \times n_{pos\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 11$
10:NP1 V CL NP ADV	Eu dou-te o livro docemente	$cl_{10} = 2 \times np \times n_{pos\_adv} \times n_{cj\_vb \wedge pr=np1\_pr} \times 6$
11:WH CL V NP	Quem te deu este livro?	$cl_{11} = n_{wh} \times np \times n_{cj\_vb \wedge pr=wh\_pr} \times 6$
12: WH CL V	Quem o viu?	$cl_{12} = n_{wh} \times n_{cj\_vb \wedge pr=wh\_pr} \times 11$
13: NP V CL WH	Tu deste-o a quem?	$cl_{13} = np \times n_{wh} \times n_{cj\_vb \wedge pr=wh\_pr} \times 11$
14: QT NP CL V	Poucas pessoas o lêem	$cl_{14} = n_{dqt} \times np \times n_{cj\_vb \wedge pr=np\_pr} \times 11$
15: QT NP1 CL V NP	Poucas pessoas te lêem livros	$cl_{15} = n_{dqt} \times 2 \times np \times n_{cj\_vb \wedge pr=np1\_pr} \times 6$
16: QT NP V CL	Algumas pessoas lêem-no	$cl_{16} = n_{ndqt} \times np \times n_{cj\_vb \wedge pr=np\_pr} \times 11$
17: QT NP1 V CL NP	Algumas pessoas lêem-te livros	$cl_{17} = n_{ndqt} \times 2 \times np \times n_{cj\_vb \wedge pr=np1\_pr} \times 6$
Complex Clauses		$n_{complex\_clauses} = ccl$
18: Clause CONJ Clause	Eu dei-te o livro, para tu o leres	$ccl = 2 \times n_{clauses} \times n\_conj$
		$n_{sentences} = n_{complex\_clauses} + n_{clauses}$

Table 1. Templates of sentences that can be parsed and produced in our design, and expected complexity. (n - number of; cj_vb - conjugated ver tense; np_pr - person of noun phrase; n/op_adv - non/operator-type adverb; n/dqt - non/downward quantifier)

5 Conclusions and Further Research

In this paper, we presented an initial computational implementation to parse and produce sentences in EP with the correct placement of pronominal clitics. The sentences handled are still limited. In the future, we would also like to explore more proclitic contexts, the morphological aspects of the clitics, and automatic creation of constructions to expand the grammar with an annotated corpus.

Acknowledgments. The research presented in this paper has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 *ESSENCE: Evolution of Shared Semantics in Computational Environments* (`http://www.essence-network.com/`).

References

1. Spencer, A., Luís, A. R.: Clitics: an introduction. Cambridge University Press (2012)
2. Silva, J., Branco, B., Gonalves, P.: Top-Performing Robust Constituency Parsing of Portuguese: Freely Available in as Many Ways as you Can Get it. LREC. (2010)
3. McDonald, R.: Discriminative learning and spanning tree algorithms for dependency parsing. Dissertation, University of Pennsylvania. (2006)
4. Bick, E.: The parsing system Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. (2000)
5. Luís, A. R., Otoguro R.: Inflectional morphology and syntax in correspondence. Morphology and Its Interfaces, 178, 97–135 (2011)
6. Bresnan. J.: Lexical Functional Syntax. Blackwell Pub. Oxford (2001)
7. Steels, L.: Design patterns in Fluid Construction Grammar. John Benjamins Amsterdam. (2011)
8. Steels, L.: Basics of Fluid Construction Grammar. Constructions and Frames (*in press*)
9. Wang, Y., Berant, J. Liang, P.: Building a semantic parser overnight. Association for Computational Linguistics (ACL). (2015)
10. Luís, A. R., Otoguro R.: Proclitic contexts in European Portuguese and their effect on clitic placement. In Proceedings of the LFG'04 Conference (2004)
11. Van Valin, R. D., La Polla, R.J.: Syntax: Structure. Meaning and Function, Cambridge University Press, Cambridge (1997)
12. Crysmann, B.: Constraint-based Coanalysis. PhD thesis, Universität des Saarlandes and DFKI Gmbh (2002)