Detecting Erroneous Uses of Complex Postpositions in an Agglutinative Language

Arantza Díaz de Ilarraza **Koldo Gojenola** **Maite Oronoz**

IXA NLP group. University of the Basque Country

jipdisaa@si.ehu.es koldo.gojenola@ehu.es maite.oronoz@ehu.es

Abstract

This work presents the development of a system that detects incorrect uses of complex postpositions in Basque, an agglutinative language. Error detection in complex postpositions is interesting because: 1) the context of detection is limited to a few words; 2) it implies the interaction of multiple levels of linguistic processing (morphology, syntax and semantics). So, the system must deal with problems ranging from tokenization and ambiguity to syntactic agreement and examination of local contexts. The evaluation was performed in order to test both incorrect uses of postpositions and also false alarms.

1 Structure of complex postpositions

Basque postpositions play a role similar to English prepositions, with the difference that they appear at the end of noun phrases or postpositional phrases. They are defined as "forms that represent grammatical relations among phrases appearing in a sentence" (Euskaltzaindia, 1994). There are two main types of postpositions in Basque: (1) a suffix appended to a lemma and, (2) a suffix followed by a lemma (main element) that can also be inflected.

(1) *etxe-**tik***
house-(from the)
from the house

(2) *etxe-**aren gain-etik***
house-(of the) top-(from the)
from the top of the house

The last type of elements has been termed as *complex postposition*. We will use this term to name the whole sequence of two words involved, and not just to refer to the second element. Complex postpositions can be described as:

(3) lemma_1 + (suffix_1 + lemma_2 + suffix_2)

In these constructions, the second lemma is fixed for each postposition, while the first lemma allows for much more variation, ranging from every noun to some specific semantic classes. The above description (3) is intended to stress (with parentheses) the fact that the combination of both suffixes with the second lemma acts as a complex case-suffix that is "appended" to the first lemma. Both suffixes present different combinations of number and case, which can agree in several ways, depending on the lemma, case or contextual factors. Table 1 shows the different variants of two complex postpositions, derived from the lemmas *bitarte* and *aurre*. For example, the lemma *bitarte* is polysemous ("means, by means of, instrument, while (temporal), between"). Multiple factors affect the correctness of a postposition, including morphological and syntactic constraints. We also discovered a number of relevant contextual factors, which are not explicitly accounted for in standard grammars.

2 The corpus

The detection of erroneous uses of complex postpositions needs first a corpus that can serve for both development and evaluation of the system. To obtain such a corpus is a labor-intensive task, to which it must be added the examination and markup of incorrect examples. The use of a big "correct" corpus will allow us to test our system *negatively*, thoroughly testing the system's behavior in respect to false alarms. We used an automatic system for detecting complex postpositions in order to get development and test data. There are two text types: Newspaper corpora (henceforth NC, 8,207,919 word-forms) that is subject to an edition process and style guides, and Learner corpora (LC, 994,658 word-forms), which come from texts written by learners of Basque and University students. These texts are more "susceptible" of containing errors.

Coling 2008: Companion volume – Posters and Demonstrations, pages 31–34
Manchester, August 2008

lemma$_2$	suffix$_1$	suffix$_2$	Examples
bitarte (noun)	*-en* (genitive)	*-z* (instrumental)	*etxearen bitartez* (by means of the house)
	-ra (alative)	*-n* (inessive, sg.)	*etxera bitartean* (while going to the house)
	-a (absolutive, sg.)	*-n* (inessive, sg.)	*ordubata bitartean* (around one o´clock)
	-ϕ (no case)	*-n* (inessive, sg.)	*meza bitartean* (while attending mass)
	-en (genitive)	*-n* (inessive, sg.)	*mendeen bitartean* (between those centuries)
	-ϕ (no case)	*-ϕ/ko* (no case/genitive)	*Lau hektarea bitarte* (in a range of four hectares)
	-ak (absolutive, pl.)	*-ϕ/ko* (no case/genitive)	*seiak bitarte* (around six o'clock)
	-ra (alative)	*-ϕ/ko* (no case/genitive)	*etxera bitarte* (in the way home)
aurre (noun)	*-ϕ/-en* (no case/genitive)	*-n/-ra/-tik/-ko* (inessive/ alative/ ablative/ genitive)	*eliza aurrean* (in front of the church)
	-tik (ablative)	*-ra* (alative)	*hemendik aurrera* (from here onwards)

Table 1. Complex postpositions for *bitarte* and *aurre*.

We decided to study those types of postpositions that appear most frequently in texts, those containing the following lemmas as their second element: *arte*, *aurre*, *bitarte*, *buruz*, and *zehar*[2]. We selected these postpositions given that they are well documented in grammar books, with detailed descriptions of their correct and incorrect uses (e.g. see Table 1 for *bitarte*), and also that they are very frequent in both types of texts.

Each kind of syntactic error occurs with very low frequency and, therefore, big corpora are needed for evaluation and testing[3]. Even if such corpora are available, to obtain naturally occurring test data, hundreds of texts should be manually examined and marked. As a result, we decided to only manually mark errors in Learners' Corpora (LC), because NC, an order of magnitude bigger than LC, is presumed to contain less errors. This implies that we will be able to measure precision[4] in both corpora, while recall[5] will only be evaluated in LC. Table 2 shows the number of sentences used for development (60% of each corpus) and test (40%). We treated LC and NC separately, as they presumably differ in the number of errors.

3 Linguistic Processing Tools

The corpus was automatically analyzed by means of several linguistic processors: a morphosyntactic analyzer (Aduriz et al., 2000), EUSTAGGER, the lemmatizer/tagger for Basque, and the Constraint Grammar parser (CG, Tapanainen, 1996) for morphological disambiguation.

Added to these, we also used other resources:

- Grammar books which describe errors in postpositions (Zubiri & Zubiri, 1995).
- Place names. Two of the selected postpositions (*arte*, *aurre*) are used in expressions that denote temporal and spatial coordinates, but their variants impose different restrictions and agreement (case, number). In order to recognize common nouns that refer to a spatial context, we made use of a new lexical resource: electronic versions of dictionaries (Sarasola, 2007; Elhuyar, 2000). 168 and 242 words were automatically acquired from each dictionary. To this, we added proper names corresponding to places.
- Animate/inanimate distinction. Regarding postpositions formed with *aurre*, Zubiri et al. (1995) point out that "typically the previous word takes the genitive case, although it can also be used without a case mark with inanimate nouns". For this reason, we used a dictionary enriched with semantic features, such as animate/inanimate, time or instrument. We selected 1,642 animate words. We also added person names and pronouns.

4 Rule design

The system will assign an error-tag to those word-forms that show the presence of an incorrect use of a postposition. We use the CG formalism (Tapanainen, 1996) for this task. CG allows

	NC		LC	
	Dev	**Test**	**Dev**	**Test**
arte	7769	5179	1209	806
aurre	8129	5420	1157	771
bitarte	3846	2564	772	514
buruz	5435	3623	560	373
zehar	1500	1000	186	126
Total	26679	17786	3884	2590
Errors			60	29

Table 2. Number of sentences in development and test sets, including the errors in LC.

[2] As each lemma has several meanings depending on each variant, we will not give their translation equivalence.

[3] We made an estimate of more than 1% of elements in general corpora being complex postpositions.

[4] Number of errors correctly identified by the system / total number of elements identified as erroneous.

[5] Number of errors correctly identified by the system / total number of real errors.

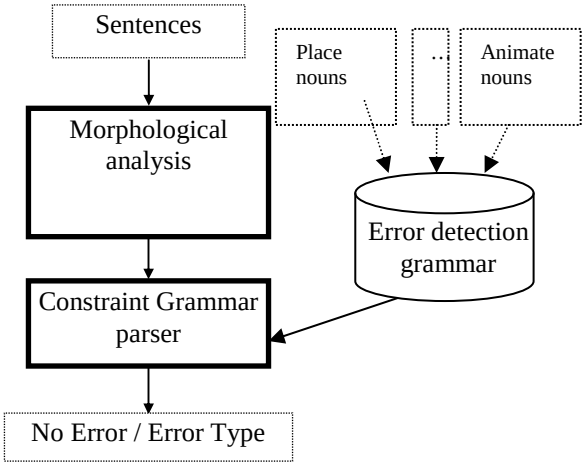

Figure 1. General architecture.

the definition of complex contextual rules in order to detect error patterns by means of mapping rules and a notation akin to regular expressions.

Fig. 1 shows a general overview of the system. Syntactic constraints are encoded by means of CG rules using morphosyntactic categories (part of speech, case, number, …). Semantic restrictions are enforced by lists of words belonging to a semantic group. All of the five postpositions have clear requirements about the combinations of case and number in the surrounding context.

Overall, the CG grammar contains 30 rules for the set of 5 postpositions. We found that although the study of authoritative grammatical descriptions was exhaustive, the grammarians' descriptions of correct and incorrect uses refer mainly to morphology and syntax. Nevertheless, we discovered empirically that most of the rules needed to be extended with several classes of semantic restrictions. Among others, distinctions were needed for animate nouns, place names, or several classes of time expressions, depending on each different variant of each postposition.

5 Evaluation

The rules were applied both to the (presumably) correct newspapers texts (NC) and to the learners' texts (LC). The actual errors in LC were marked in advance but not in NC, which means that recall can only be evaluated in LC. Table 3 shows the main results including all the selected five postpositions. The LC corpus contains 60 and 29 error instances in development and test corpus, respectively. If we concentrate on precision, Table 4 shows the overall precision results for the total of errors detected in the test corpora. When we consider the whole set of postpositions precision is 50.5%, giving 42 false alarms out of 85 detected elements. We performed an analysis of false alarms which showed several causes:

	NC		LC	
	Dev	**Test**	**Dev**	**Test**
Sentences	26679	17786	3884	2590
Errors	-	-	60	29
Undetected	-	-	10	10
Detected	30	24	50	19
False alarms	45	33	2	9
Recall	-	-	83%	65%
Precision	40%	42%	96%	67%

Table 3. Evaluation results.

Postpositions	Precision
arte, aurre, bitarte, buruz, zehar	**50.5%** (43/85)
aurre, bitarte, buruz, zehar	**70.4%** (31/44)
bitarte, buruz, zehar	**78.3%** (29/37)

Table 4. Precision for the test sets (NC + LC).

- Morphological ambiguity (43% of alarms).
- Semantic ambiguity (28%). We included sets of context words to identify the correct senses, but it still causes many false alarms.
- Syntactic ambiguity (22%). The false alarms are mostly concerned with coordination.
- Tokenization errors (7%).

As most of the false alarms came from postpositions formed with *arte*, the most ambiguous one, we counted the errors when dealing only with the other four postpositions, giving a better precision (70.4%, second row in Table 4), although detecting less true errors. If the system only deals with three postpositions (third row in Table 4), then precision reaches 78.3%. Johannessen et al. (2002) note that the acceptable number of false alarms in a grammar checker should not exceed 30%, that is, at least 70% of all alarms had to report true errors. Our experiments show that our system performs within that limit, albeit restricting its application to the most "profitable" postpositions. Although the number of rules varies widely (from 15 rules for *arte* to 2 rules in the case of *zehar*) their effectiveness greatly depends on the complexity and ambiguity of the contextual factors. For that reason, *arte* presents the worst precision results even when it contains by far the biggest set of detection rules. On the other hand, *zehar*, with 2 rules, presents the best precision, due to its limited ambiguity. So, to deal with the full set postpositions (several works estimate more than 150), it will be more profitable to make a preliminary study on ambiguities and variants for each postposition.

6 Related work

Kukich (1992) surveys the state of the art in syntactic error detection. She estimates that a proportion of all the errors varying between 25% and over 50% are valid words. Atwell and Elliott

(1987) concluded that 55% of them are local syntactic errors (detectable by an examination of the local syntactic context), 18% are due to global syntactic errors (which need a full parse of the sentence), and 27% are semantic errors. Regarding their treatment, there have been proposals ranging from error patterns (Kukich 1992; Golding and Schabes 1996), in the form of hand-coded rules or automatically learned ones, to systems that integrate syntactic analysis.

(Chodorow et al., 2007) present a system for detecting errors in English prepositions using machine learning. Although both English prepositions and Basque postpositions have in some part relation with semantic features, Basque postpositions are, in our opinion, qualitatively more complex, as they are distributed across two words, and they also show different kinds of syntactic agreement in case and number, together with a high number of variants. This is the main reason why we chose a knowledge-based method.

7 Conclusions

We have presented a system for the detection of errors in complex postpositions in Basque. Although at first glance it could seem that postpositions imply the examination of two consecutive words, a posterior analysis showed that they offer rich and varied contexts of application, requiring the inspection of several context words, albeit not enough to need a full syntactic or semantic analysis of sentences. The system uses a varied set of linguistic resources, ranging from morphological analysis to specialized lexical resources. As the detection of these errors implies a detailed and expert linguistic knowledge, the system uses a purely knowledge-based approach.

A considerable effort has been invested in the compilation of a corpus that provides a testbed for the system, which should be representative enough as to predict the behaviour of the system in an environment of a grammar checker. For that reason, we have tried to put a real emphasis on avoiding false alarms, that is, treating also lots of correct instances. The results show that good precision can be obtained. Regarding recall, our experiments do not allow to make an estimation, as the NC test corpora is too big to perform a detailed examination. However, the LC corpora can give us an upper bound of 65% (see Table 3).

This work also shows that the use of purely morphosyntactic information is not enough for the detection of errors in postpositions. For that reason we were forced to also include several types of semantic features into the system. On the other hand, the process of automatic error detection has also helped us to explore new sets of semantic distinctions. So, the process of error detection has helped us to organize concepts into sets of semantically related elements, and can serve to make explicit types of knowledge that can be used to enrich other linguistic resources.

We can conclude saying that descriptive linguistics could benefit from error diagnosis and detection, as this could help to deeply understand the linguistic descriptions of postpositions, which are done at the moment mainly by means of morphosyntactic information, insufficient to give an account of the involved phenomena.

Acknowledgements

This research is supported by the University of the Basque Country (GIU05/52) and the Basque Government (ANHITZ project, IE06-185).

References

Aduriz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J., Artola X., Gojenola K., Sarasola K. 2000. A Word-grammar based morphological analyzer for agglutinative languages. *COLING-00*.

Atwell E., Elliott S. (1987) Dealing with Ill-Formed English Text. *In The Computational Analysis of English: a Corpus-Based Approach*. Longman.

Chodorow M., Tetreault J. and Han N. 2007. Detection of Grammatical Errors Involving Prepositions. *4th ACL-SIGSEM Workshop on Prepositions*.

Díaz de Ilarraza A., Gojenola K., Oronoz M. 2008. Detecting Erroneous Uses of Complex Postpositions in an Agglutinative Language. Internal report (extended version). (https://ixa.si.ehu.es/Ixa/Argitalpenak)

Elhuyar. 2000. *Modern Basque Dictionary*. Elkar.

Euskaltzaindia. 1994. *Basque Grammar: First Steps* (in Basque). Euskaltzaindia.

Golding A. and Schabes. Y. (1996) Combining trigram-based and feature-based methods for context-sensitive spelling correction. *ACL 1996*.

Johannessen J.B., Hagen K., and Lane P. 2002. The performance of a grammar checker with deviant language input. *Proceedings of* COLING, Taiwan.

Kukich K. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*.

Tapanainen P. 1996. *The Constraint Grammar parser CG-2*. Publications of the Univ. of Helsinki, 27.

Sarasola, Ibon. 2007. *Basque Dictionary* (in Basque). Donostia : Elkar, L.G. ISBN 978-84-9783-258-8.

Zubiri I. and Zubiri E. 1995. E. *Euskal Gramatika Osoa* (in Basque). Didaktiker, Bilbo.