Procesamiento del lenguaje Natural, nº 41 (2008), pp. 13-19 recibido 7-05-2008; aceptado 16-06-2008

Analysis of Noun-Noun sequences[1]: a rule based approach

Análisis de secuencias N-N: un enfoque con gramáticas basadas en reglas

Jose Mari Arriola
UPV/EHU-Basque Philology Department
School of Economy and Business
Oñati Plaza 1, 20018 Donostia
josemaria.arriola@ehu.es

Juan Carlos Odriozola
UPV/EHU-Basque Philology Department
Science and Technology Faculty
Leioa 48940 (Bizkaia)
juancarlos.odriozola@ehu.es

Abstract: This paper reports on work in progress to improve shallow parsing for Basque. The practical goal of our work is to enrich the information of the shallow parser with linguistic information for analyzing sequences containing an N that instantiates a kind of quantification of the other nominal constituent, by means of some different syntactical structures.

Keywords: shallow parsing, noun phrase chunking.

Resumen: El artículo presenta el trabajo para mejorar el parser superficial del euskara. El objetivo práctico del mismo, consiste en enriquecer dicho parser con la información lingüística pertinente para analizar secuencias que contienen un elemento nominal que instancia por medio de diversas estructuras sintácticas algún tipo de cuantificación de un segundo N.

Palabras clave: parsing superficial, chunks de sintagmas nominales.

1 Introduction

The general framework of the research work and implementation reported here is the syntactic-processing system of Basque (see 1). We are working on a robust parsing scheme that provides syntactic annotation in an incremental fashion: once textual input has been tokenized, morphologically analyzed and disambiguated, syntactic annotation is added in two distinct stages of processing. First, a chunk parser provides a partial constituent analysis. In a second stage, the chunked input is further annotated by dependency links.

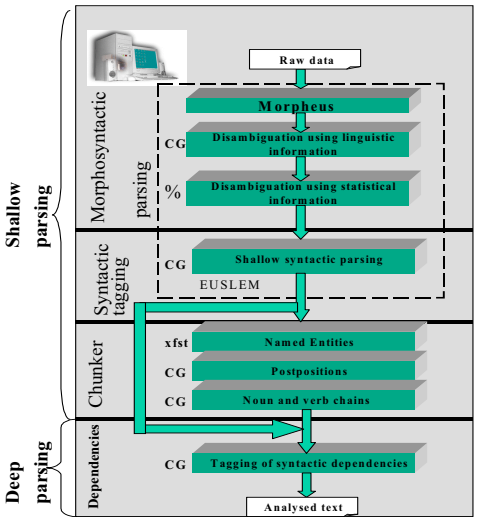

Figure 1: General framework

[1] This research is supported by grants no. HUM2004-05658-C02-01, UPV 1/UPV 00113.310-H-15921/2004 and EHU06/16, HUM2004-05658-C02-01 and EHU06/16. Besides, acknowledgments to the support of the Government of the Basque Country to IXA group.

This paper addresses one specific subtask in the overall parsing scheme: improvement of the shallow analysis of a subclass of nouns involved in two specific syntactic structures that bear a kind of quantification. These syntactic structures require specific syntactic rules that qualitatively improve the NP chunking step, in that these rules are applied only to a subclass of nouns that are close to quantifiers. We will therefore be concerned with a lexical/functional class that is not as wide as the entire noun class. Nevertheless this noun class is wider than the close functional class of quantifiers. We will describe in detail the linguistic information that can be applied for dealing with quantifying N-N sequences instantiated in both quantifying compounds and measure phrases.

Noun sequences have certain characteristics that hinder automatic interpretation. Firstly, the creation of noun sequences is highly productive in Basque; it is not possible to store all the noun sequences that will be encountered while processing text. Basque N-N sequences could be said to pattern with English counterparts, and hence, most of the N-N sequences can be interpreted or processed as restriction readings of the second N, which is the head in the constructions of both languages. Indeed, an *itsasgizon* lit. 'seaman' is a kind of *gizon* 'man'. Secondly, their interpretation is not always recoverable from syntactic or morphological analysis. As is well known that most of these sequences bear a non-compositional meaning, and *baserri* lit. 'wood town' means 'farm'. Crucially, some of the Basque N-N sequences contain nominal heads that rarely appear in English counterparts. These Ns have not suffered a lexicalisation. On the contrary, they are similar to quantifiers, in that they force quantifying readings, i.e., not lexical but functional features seem to appear, for instance: *esne pixka bat* lit. 'milk bit a' or *mutil mordo bat* lit 'boy lot a'.

Thirdly, some of the right Ns in the sequence supposed to be a compound (*esne-botila bat* lit. 'milk bottle one' bear two kinds of readings (lit. 'to drink/to break one milk bottle') but they also appear in a totally different structure of the type *botila bat esne* lit. 'bottle one milk', which takes a single content reading. In all the other nouns outside this subclass (cif. *itsasgizon* 'seaman', *baserri* 'farm') the first structure is not available and in the second structure they bear a unique reading corresponding to themselves.

In this work, *esne-botila bat, botila bat esne* will all be called quantifying N-N sequences, and *botila, pixka* and *mordo* will be quantifying nouns.

The practical goal of our work is to enrich the information in mapping rules of the chunk parser with linguistic information dealing with quantifying N-N sequences.

This approach could be useful for additional processing (deep syntax) or for end applications (data mining, IR, etc.).

The rest of this paper is organised as follows: Section 2 describes the previous work on chunking; Section 3 describes the linguistic knowledge needed for improving the NP chunking; Section 4 is devoted to specifying how to apply the linguistic information in a rule-based approach; Section 5 shows the experiments performed. Finally, some conclusions are outlined in Section 6.

2 Previous work

There is an extensive bibliography on the processing of nominal sequences, based on (sub)categorization features of the constituents (Barker, 1998), (Takeuchi et al., 2001), (Flickinger & Bond, 2003). Ngai and Yarowsky (2000), on the other hand, present a comprehensive empirical comparison between two approaches for developing a base noun phrase chunker: human rule writing and active learning using interactive real-time human annotation.

In this section we describe the main steps followed in our shallow syntactic analysis of the corpus. The main base in the analysis of the corpus is the morphological analyser (Alegria et al. 1996) and the disambiguation grammar (Aduriz et al. 2000). Using eliminative linguistic rules or constraints, contextually illegitimate alternative analyses are discarded by means of Constraint Grammar (CG) (Karlsson et al. 1995). This gives us almost fully disambiguated sentences, with one interpretation per word-form and one syntactic-tag label. But there are word-forms that are still morphologically and syntactically ambiguous. At this point we are aware that shallow syntax is the best approach for robust syntactic parsing. As our base, we took the surface

oriented syntactic tags in order to analyze noun chains and verb chains. Despite the remaining ambiguity and errors, the identification of various kinds of chunks is reasonably straightforward. For this purpose we based our work on the syntactic function tags designed for Basque (Aduriz et al. 1997). We can divide these tags into three types: main function syntactic tags, modifier function syntactic tags and verb function syntactic tags. This distinction of the syntactic functions was essential for the CG-style subgrammars that contain mapping rules.

The first version of the shallow grammar was applied over a sample of 300 sentences (extracted at random from Euskal Hiztegia). This was manually checked and the proportion of sentences that had the noun and verb chains tags correctly assigned was 75% (Arriola et al., 1999). This grammar has recently been improved (Aranzabe et al. 2004).

At this stage we are concerned with noun chunks: those phrase units headed by a noun.

For this reason, we will explain the subgrammar for noun chunks. The assumption is that any word having a modifier function tag is linked to some word with a main syntactic function tag. Moreover, a word with a main syntactic function tag can by itself constitute a chunk or phrase unit.

The syntactic representation of noun chunks was based on the following syntactic tags:

- @ID>/ @<ID: pre/postmodifying determiner.
- @IA>/@<IA: pre/postmodifying adjective.
- @IZLG>/@<IZLG: pre/postmodifying noun complement.
- @KM>: modifier[2] of the element containing the case and determination. This is the element with a main syntactic function tag.

Using this assumption we established three tags to detect this kind of chunk:

- %NCH: this tag is attached to words with main syntactic function tags that constitute a chunk by themselves.
- %INIT_NCH: this tag is attached to words with main syntactic function tags that are linked to other words with modifier syntactic function tags and constitute the initial element of a phrase unit.
- %FIN_VCH: this tag is attached to words with main syntactic function tags that are linked to other words with modifier syntactic function tags and constitute the end of a chunk.

The aim of this subgrammar is to attach to each word-form one of those three tags in order to delimit the noun chunks. They make explicit the linking relations expressed by the syntactic functions. In Fig. 2 there is an example[3] that shows how the analysis of the chunker is equivalent to the analysis of a sentence into phrases[4]:

```
"<Hipoteka-kreditu>" <INIT_CAP>"                          mortage
    "hipoteka-kreditu" N  @ KM>     % INIT_NCH
"<zati>"                                                  piece
    "zati" N              @ KM>
"<handi>"                                                 big
    "desobedientzia" N    @ <IA
"<bat>"                                                   one
    "bat" ADJ             @ <ID     % FIN_NCH

"<ordainatzeke>"                                          unpayed
    "ordaindu" V       @-FMAINV   %INIT_VCH
"<dugu>"
    "*edun" AUXV  @+FAUXV         %FIN_VCH        we have
"<$.>"<PUNCT_PUNCT>
```

Fig. 2. Analysis of chains. English translation on the right.

3 Linguistic Knowledge

The linguistic information summarized in this section is based on the linguistic data provided by Odriozola (2006, 2007, 2008). Henceforth, we shall talk about both "quantifying nouns" and "quantifying sequences" as long as a measure noun and a measured noun are involved in a Basque syntactical structure.

[2] Basque is a head-final language provided with postpositions, so that @<KM is not needed.

[3] Each syntactic function tag is prefixed by "@" in contradistinction to other types of tags. Some tags include an angle bracket, "<" or ">". The angle bracket indicates the direction where the head of the word is to be found.

[4] The syntactic structure of the noun chunk in other terms: [[[[hipoteka-kreditu] zati]Iz handi]NP bat D]DS.

Some of the quantifying nouns appear in both constructions taken as compounds and in measure phrases. Some others appear only in one of the two constructions. The distribution of the several nominal constituents will be taken into account here.

Following Solé (2002) and Odriozola (2008), we assume that there is a kind of Basque (measure) noun that individualizes mass nouns by means of the following syntactical patterns.

Content nouns are involved in both measure phrases headed by the mass noun (1) and structures headed by the content noun itself that has usually been taken as compounds(2):

(1) hiru botila esne (gozo)
three bottle milk (sweet)
'three bottles of (sweet) milk'

(2) a Hiru esne (*gozo) botila apurtu ditugu
three milk sweet bottle broken AUX.
'We broke three bottles of (sweet) milk'

b Hiru esne (*gozo) botila edan ditugu
three milk sweet bottle drunk AUX.
'We drank three bottles of (sweet) milk'

It should be remarked that the measure phrase in (1) actually bears two phrases, [hiru botila] and [esne (gozo)]. In any case the double readings are common in human languages (Castillo 2001). This is not so in the Basque second option. Furthermore, the so-called compound may bear either a container reading (2a) or a content reading (2b).

Unit nouns are involved in the measure phrases described above (2a). They rarely appear in quantifier compounds (2b)

(3) a bi litro esne
two liter milk
'two liters of milk'

b %bi esne-litro
two milk liter

Some nouns are claimed to be grammaticalized to (a complex) quantifier, since they can only appear with the quantifyer/determiner *bat* 'one/a'

(4) a *esne pittina
milk bit-DET

b esne pittin bat
milk bit one
'a little bit of milk'

It is worth remarking that this kind of (quantifier) compound-like constructions can only take a conceptual reading corresponding to the left constituent presumed not to be the head.

Following Solé (2002) we assume that there are some Basque collectivizing nouns that head (quantifier) compounds similar to those headed by the individualizing nouns. Needless to say, the left constituent here is a countable noun, and the reading often corresponds to a reading related to the non-head constituent

(5) a mutil mordo bat
boy lot one
'a lot of boys'

b mutil mordoa
boy lot-**DET**

It should be observed that this kind of quantifying noun is never totally grammaticalized and they always accepts the attached determiner *–a*, as standard nouns do. On the other hand, some such nouns may take a reading that is somewhat independent of the left constituent and may even force a singular agreement in the verb:

(6) mutil mordoa etorri dira
boy lot-det come-PERF AUX-PL
'A lot of boys came'

(7) mutil taldeak ondo jokatu du
boy team-DET well play-PERF AUX-SING

Nouns like parte *zati* `piece´ and *tarte* `interval´ express a non-specific part or a whole that can be mass as in *ogi zati* lit 'bread piece' or something that is subcategorized as a mass

like *opil zati* lit 'muffin piece'. Sequences of this type rarely allow more than two elements in Basque. However**,** the language allows these kind of left components when the right component is either a part noun or a collective noun. The ability of both part and collective nouns to allow a noun phrase to the left is an evidence of the non (clear) compound nature of Basque quantifying N-N sequences.

4 Rule based grammar

The linguistic information described above has been implemented by means of CG style mapping rules for adequately analyzing the cases established before.

In order to maintain coherence in quantifying relation when the element carrying the quantifying information is a noun, we decided to include new syntactic function tags: @<NQ which stands for a noun quantifier that modifies the noun to the left; and @NQ> for a noun quantifier that modifies a noun to the right.

In the case of the quantifying compounds, the @<NQ tag will be attached to the second element in the construction as a quantifier of the first nominal element. We deal with particular N-N sequences where the second nominal element supposed to be the head[5] of the construction somehow instantiates a quantification of the first nominal element, so the reading actually corresponds to the non-head constituent. This function will be attached to those quantifying nouns that have been detected, for instance: part nouns (zati), collective nouns (mordo) or complex determiners (pittin bat). Here are some examples:

i)[**Hipoteka-kreditu** @KM> **zati** @<NQ bat @OBJ]

ii))[*Mutil* @KM> **mordo** @<NQ *bat* @OBJ @SUBJ]

iii) [*Esne* @KM> **pittin** @<NQ *bat* @OBJ @SUBJ]

The @NQ> tag will be attached to the first element of the construction as a quantifier of the second nominal element. This function will be attached to those quantifying nouns that have been detected, for instance: content nouns that can also be involved in measure phrases (*botila*) or unit nouns (*litro*). For instance:

iv) [**Botila** @NQ> *bat* @ID> *esne* @OBJ @SUBJ]

v) [*Bi* @ID> **litro** @NQ> *esne* @OBJ @SUBJ]

The analysis introduces some idiosyncratic constructions, the *noun quantifying rules*, which links together syntax and morphology. Combined with existing rules, the new rules accounts for the both the distributional and agreement idiosyncrasies.

5 Experiments

We divide the available data into a train and test set, trained the CG grammar on the train set and compared the results on the test set. These rules were formulated, implemented and tested using selected examples from Twentieth Century Basque Corpus[6].

In addition, we have taken a sample of 1, 737 noun chunks from EPEC (corpus of standard written Basque that has been manually tagged at different levels (morphology, surface syntax, phrases).

Our results were taken after applying the mapping rules to the output of the noun chunker. The parser labels the selected examples by attaching every quantifying noun to the noun head by means of the corresponding quantifying tag function (@<QN, @QN>). When we examine the noun chunks function tags which do not distinguish between quantifying nouns (@<QN, @QN>) and case-marker modifier nouns (@KM>), these differences do not affect the noun chunk segment. An example is the [Hipoteka-kreditu zati bat] mentioned above. The operation of text chunking, consisting of dividing a text into syntactically correlated parts of words, has not changed.

[5] N1-N2 sequences are described in Basque are of two types: Either N1 syntactically and semantically depends on N2 or dependency cannot be checked. Unlike romance languages such as Spanish, Basque rarely produces left-headed N-N sequences.

[6] http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html.

However, since there were no syntactic function tags to distinguish quantifying nouns, the parser did not know which the head was: "hipoteka-kreditu" or "zati". The nouns labelled with the noun quantifying tags are similar to quantifiers, in that they force quantifying readings. In fact, *zati* cannot be the lexical head. From a more formal point of view, *zati* would be the functional head of the head construction, whereas the other nominal would be the lexical head.

We have evaluated the precision (correctly tagged quantifying nouns/total number of quantifying nouns) and recall (relevant tagged quantifying nouns/actual quantifying nouns in the corpus). For quantifying nouns tag precision and recall were 93% and 100% respectively. The errors are due to the remaining ambiguity in the morphosyntactic analysis.

At the same time, we performed an experiment to question the idea that the noun quantifying tags give better results in determining the head of N-N sequences. We tried a CG style grammar to attach the head tag (&Head) concluding that in most cases this grammar gives good results. However, there are also many examples where the shallow syntactic information we are using is not sufficient to determine the lexical head. For instance, in the case of container nouns that can appear in both quantifying compounds and measure phrases we have two readings (content and container). In these cases we need information about the verb. In the same sense, in the case of collective nouns like *talde* we have two interpretations, for example:

1. neska **taldea** etorri zait ("**many** girls have come to me"

2. a. zuzendari **talde** bat aukeratu dute

'They have chosen some managers'

b. zuzendari-**taldea** aukeratu dute

'They have chosen the management team'

3. c. Zuzendaritza-**taldea** aukeratu dute

'They have chosen the direction team'

Both (1) and (2a) bear a quantified reading of the left constituent. (2b) seems to not to bear a quantifying reading and talde has a specific reading. Finally, (3c) takes a reading that is clearly not-quantifying.

We consider these results satisfactory as a first approach, even more so if we take into account the fact that the work is still in progress and also that, in some cases there is a lack of sufficient data in our corpus.

6 Conclusions and future work

The rules have been implemented and tested in the CG grammar, a broad-coverage grammar of Basque. Our analysis supports the position that broad-coverage grammars will necessarily contain both highly schematic and highly idiosyncratic rules. Our approach to improving parsing is to modify the syntactic tagset and to add mapping rules that are used for attaching those new syntactic tags to quantifying nouns.

The results are satisfactory in the case of tagging quantifying nouns with the new syntactic function tags (@<QN, @QN>) and we achieve considerable improvement, when head labelling is performed on noun chunks.

Apart from the evaluation of the results obtained by the mapping rules for detecting noun-modifier structures, we wanted highlight the benefits of using the enhanced syntactic analysis for detecting the lexical heads of NPs. The previous shallow analysis of NPs did not include the noun-modifier function so that in the case of those specific structures there is more than one candidate head. With the near perfect rates of recall and precision obtained in the analysis of those noun-modifier structures the automated extraction of term candidates from text will be improved. Unfortunately, we have not yet been able to use the enhanced version in real situations of terminology work, so we cannot give exact figures.

Besides, as already mentioned before, the creation of noun sequences is highly productive in Basque; it is not possible to store all the noun sequences that will be encountered while processing. For this reason and for future work we plan to write some grammar rules for detecting previously undetected quantifying nouns.

Indeed, as far as these nouns are involved in quantification, we could assume that they belong to a subclass that is closed, although it must be large.

The information associated with these grammar rules is as follows:

- Lexical information: mass nouns. Combined with the syntactic information on numerals and the plural overt agreement in the finite verb.
- Morphological information: the particular morphology as indicator is in bold: `goilara**kada** bat azukre´ lit spoonful one sugar; ´bi opil hiru**ren**´ lit one third of the muffin or specific collective nouns: bi**kote** pair; hiru**kote** trio.
- In hyponim/hypernym relationships, the class of beings expressed by one of the nouns is a subclass of that expressed by the other noun. This article is concerned with meronymic information, where one of the nouns expresses a part of that expressed by other noun. We accept that both individualizing nouns and countable are to be collectivized bear a meronymic relation with mass nouns and collective nouns.

Finally, we wanted to emphasize the benefits of linguistically sound methods and formalisms as the core of the linguistic processors.

Bibliography

Aduriz I., Arriola J., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Euskararako murriztapen-gramatika: mapaketak, erregela morfosintaktikoak eta sintaktikoak, UPV/EHU/LSI/TR12-2000

Aduriz I., Arriola J., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. 1997 Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism Proceedings of Recent Advances in NLP (RANLP97), 282-288. Tzigov Chark, Bulgary.

Aranzabe M., Arriola J.M., Díaz de Ilarraza 2004. Towards a Dependency Parser of Basque. Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland.

Barker, K. 1998. A trainable bracketer for noun modifiers. Advances in artificial intelligence, vol. 1418.

Castillo, J.C. 2001. Thematic Relations between Nouns. Doctoral Dissertation, University of Maryland.

Flickinger, D. & Bond F., 2003. A Two-Rule Analysis of Measure Noun Phrases. Proceedings of the HPSG03 Conference, Michigan State University, East Lansing, ed. Stefan Müller.

Karlsson F., Voutilainen A., Heikkila J., Anttila A. 1995. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Kiyoko U., Koichi T., Masaharu Y., Kyo K., Teruo K. 2001. A Study of Grammatical Categories Based on Grammatical Features for Analysis of Compound Nouns in Specialized Field. Mathematical Linguistics, vol. 23, nº 1.

Ngai, G. and Yarowsky D. 2000. Rule writing or annotation: cost-efficience resource usage for noun phrase chunking. Proceedings of 38th Annual Meeting of the Association of Computational Linguistics, 117-125,Hong-Kong.

Odriozola, J.C., 2007. '(Basque) natural phrases for artificial languages. Andolin Gogoan: Essays in Honour of Prfo Eguzkitza: 707-724.

Odriozola, J. C. 2007. Measure phrases in Basque. Lakarra & José Ignacio Hualde (eds.), Studies in Basque and Historical Linguistics in memory of R. L. Trask. Supplements of International Journal of Basque Linguistics and Philology, 40 (1-2): 739-762.

Odriozola, J.C., 2008. 'Quantifier Compounds' X.Artiagoitia & J.Lakarra (eds.). Goenagarentzako omenaldia: 503-518 (in press).

Solé, E., 2002. 'Els noms col.lectius Catalans. Descripció I reconeixement'. Doctoral Dissertation. Unibersitat Pompeu Fabra.