Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 277-281 recibido 03-05-2011 aceptado 22-05-2011

Handling Reduplication in Basque: A Problem for Spell Checking

La reduplicación en euskera: un problema para los correctores ortográficos.

Dorota Krajewska y **Tamara Hernández Godoy**
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Paseo de la Universidad, 5. 01006 Vitoria-Gasteiz
dorota.krajewska@gmail.com, tamara.hernandez.godoy@gmail.com

Resumen: La reduplicación (la repetición de uno o de parte de un lexema) se trata como repetición y es ignorada o marcada como errónea en los correctores ortográficos existentes. Para la mayoría de las lenguas, ésta es una estrategia válida, sin embargo, el euskera es diferente en este sentido como muestran variados ejemplos de repeticiones lícitas de secuencias parciales o completas. Parece que este tema ha sido ignorado en las aplicaciones computacionales existentes. En este artículo, ofrecemos una descripción del fenómeno y presentamos un prototipo para incorporarlo en un corrector ortográfico que sería capaz de manejar la reduplicación mejor que los sistemas existentes.

Palabras clave: euskera, vasco, reduplicación, corrector ortográfico

Abstract: In spell-checkers, reduplication (repetition of a word, or a part of it) is often subsumed under repetition and is either ignored or treated as erroneous. For most languages it is a valid strategy, however Basque is different in this respect as it exhibits several instances of valid partial or complete repetition of phrases. This issue appears to have been mostly ignored in existing computational applications. In this paper we provide a linguistic description of the phenomenon and present a prototype to be integrated in a complete spelling and grammar checker that would be capable of handling reduplication better than existing systems.

Keywords: Basque, reduplication, spell checker

1 Introduction and outline

In many languages, if there is an exact repetition of a word in a text, it is most probably a user mistake. An exception are homographs. However, some languages' grammars make extensive use of repetition (e.g. as a device to intensify meaning) and in this case the issue should not be ignored by spell-checkers. Basque is an example of such a language. We have identified several types of reduplication, which are not correctly handled by existing spell checkers.

Currently there are two well known spell-checkers for Basque. One is Xuxen, developed by the IXA group (Agirre et al., 1992, http://www.xuxen.com), and the other is the Microsoft Word spell-checker. Neither of them deals correctly with the problem of reduplicated words. Thus, the objective of our project was to build a prototype that would handle the problem better than existing applications.

In order to gather data, we used a corpus of Basque literary texts from the 20th century and the second half of the 19th century (approximately 750,000 words) available for download from *Klasikoen gordailua* (http://klasikoak.armiarma.com). The corpus contains around 19,000 tokens of reduplication. This demonstrates that reduplication is frequent in Basque and should not be ignored by spell checkers. Moreover, it is a productive morphological process in the sense that no fixed list can cover all possible cases.

In section 1 we provide a linguistic description of the phenomena in question, a classification of different types of reduplication and the frequency of each type in our corpus. We also present the current spelling rules our application implements. In section 2 we describe how the existing spell-checkers deal with the problem. Section 3 presents our application and the initial results obtained with it.

2 Reduplication—linguistic description

Under the label of reduplication, we subsume several phenomena which are quite different in nature, although they have similar surface forms, i.e. the same lexeme (or its part) repeated. We distinguish between three types of reduplication: morphological, syntactic and lexical:

- Morphological reduplication: "(...) a type of word formation (in the broad sense, including both derivation and inflection) in which the phonological form of an affix is determined in whole or in part by the phonological form of the base to which it attaches." (Wiltshire and Marantz, 2000, p. 557).
- Syntactic reduplication: the repetition is a consequence of syntax.
- Lexical reduplication: the repetition is a consequence of lexical phenomena (most importantly homographs).

2.1 Morphological reduplication

2.1.1 Total reduplication

Total reduplication consists of repeating the whole word. In Basque it is used to intensify the meaning (e.g. *bero* 'hot', *bero-bero* 'very hot'), or to express iterativity (e.g. *lerro* 'line' and *lerro-lerro* 'in lines, in order') or distributive meaning *(banan-banan* 'one by one'). The base might be an uninflected or inflected item. It is most commonly an adjective or an adverb, but these are not the only possibilities:

- Adverbs: *emeki-emeki* 'gently', *astiro-astiro* 'slowly', *ozta-ozta* 'with great difficulty', *beti-beti* 'always';
- Adverbs/adjectives: *gorri-gorri* 'very red', *bero-bero* 'very hot';
- Nouns: *lerro-lerro* 'line by line', *patxadaz-patxadaz* 'with calm, calmly';
- Inflected nouns: *egunero-egunero* 'everyday';
- Verbal roots: *neka-neka* 'tired', *asper-asper* 'bored';
- Inflected partciples: *berotuz-berotuz* 'warming';
- Numerals: *banan-banan* 'one by one';
- Onomatopoeias: *zapla-zapla* 'slap', *mauka-mauka* 'bark';
- Pronouns: *neure-neure* 'my', *hortxe-hortxe* 'there';
- Conjunctions: *baina-baina* 'but but'.

Some, but not all, of these expressions are highly lexicalized (i.e. they function as fixed phrases). However, on the whole, reduplication is a productive morphological process in Basque. It is particularly so with adjectives and adverbs (and more productive with adjectives than with adverbs according to Hualde (2003, p. 360) . As a consequence, it is virtually impossible to create a closed and exhaustive list of such compounds.

Most compounds of this type can take inflectional endings, according to the part of speech they belong to. For instance, *gorri-gorri* 'very red' inflects as any other adjective: *gorri-gorri-a* 'the very red one', *gorri-gorri-a-rekin* 'with the very red one', etc. Possible exceptions are the comparative and superlative forms, which are infrequent with reduplicated adjectives and adverbs (although we found some examples in the corpus, e.g. *urruti-urrutiago* 'farther', *handi-handiena* 'the very biggest', *hurbil-hurbilena* 'closest'). We refer to this subtype of reduplication as 'second word inflected reduplication'. We will use 'total reduplication' to refer to cases in which both words have exactly the same form (i.e. neither of the words carries inflectional endings).

2.1.2 Sound-symbolic reduplication

In the sound-symbolic reduplication, the stem is repeated with the first consonant replaced by *m-* (if the word starts with a vowel, *m-* is added), e.g. *handi-mandi* 'tycoon', *nahas-mahas* 'confusion', *duda-muda* 'doubt', *isilka-misilka* 'silently', *saltsa-maltsa* 'mess'. Apart from these examples, which are frequent and arguably lexicalized, it is possible to find others that prove productivity: *teologia-meologia* from *teologia* 'theology' or *itzulpen-mitzulpen* from *itzulpen* 'translation'. The second word is not an independent lexical item (e.g. *mandi* cannot be used as an adjective: *gizon handia* 'big man', but it is impossible to say *gizon mandia*).

2.2 Syntactic reduplication

2.2.1 First word inflected

In this type the first word is inflected. Most common inflectional endings are instrumental indefinite case (*-z*, e.g *etxez etxe* 'from house to house', *etxe* 'house') and the partitive (*-rik*, e.g. *kalerik kale* 'from street to street', *kale* 'street'). Less frequent are the inessive (*-(e)an*, e.g. *goizean goiz* 'early in the morning', *goiz* 'early' or 'morning') and genitive (*-(r)en*, e.g. *pozaren poz* 'very happy/happily', *poz* 'happiness'). A similar case are expressions in which the first noun is plural and the second is in the indefinite form, such as *arazoak arazo* 'despite the problems', literally 'problems problem'. Although first word inflected reduplication is a syntactic rather than a morphological process (Hualde, 2003, p. 360), we decided to treat it as reduplication, because of the surface similarity with reduplication, which may confuse speakers.

2.2.2 Repetition of verbs

In Basque it is possible to find repeated verbs (more precisely, participles). The meaning is the same as the Spanish construction with an infinitive and a finite verb, as the following example shows:

(1) **Ikusi, ikusi** dut, baina ez dut gogoratzen
'Ver lo he visto, pero no me acuerdo.'
'I saw it, but I don't remember'

2.3 Lexical reduplication

2.3.1 *Egin egin*

A special case is *egin egin*, as in *egin egin dut* which might mean 'hacer he hecho', 'I did' (i.e. an example of repetition of verbs) or else the second *egin* might be a verb focalizer (*egin* apart from being a lexical verb, is also used as a focalizer):

(2) gauzak **egin egin** behar ditugu eta ez desegin.
'We have to DO things, not undo them.'

2.3.2 Homographs

Another case are homographs, e.g. *eta ETA* ('and ETA', frequently found in the press), *eta eta* (the first *eta* expresses cause, the second is a conjunction). Also possible with some Basque names (e.g. *ekaitz* 'storm' and a first name).

Reduplication type	Types	Tokens
Sound-symbolic	130	362
First word inflected	1,302	4,364
Second word inflected	2,181	8,401
Total reduplication and lexical reduplication	1,127	5,680

Table 1: Occurrences of different types of reduplication in our corpus.

2.4 Reduplication types in our corpus

Table 1 presents the types of reduplication in our corpus. The most frequent reduplication type in our corpus is the second word inflected reduplication (however, it should be taken into account that our corpus is not lemmatized, so reduplications with different inflectional endings are treated as separate types). The second most frequent type is the total reduplication. The least common type is the sound symbolic reduplication.

The data shows that reduplication is indeed frequently used in Basque. Some expressions have many occurrences (e.g. *emeki-emeki* 'softly': 235 occurrences, *ixil-ixila* 'quiet': 227, *mendez mende* 'down through the centuries': 145). However, what is probably more important for our project is that many examples occur only once in the whole corpus. For instance, it is the case with approximately half of the types of total reduplication. It proves productivity of the process and therefore makes it necessary to look for a strategy that would not rely on a closed list of expressions in order to deal correctly with the phenomenon.

2.5 The spelling rules

We decided to implement Euskaltzaindia's spelling rules in our application (Euskaltzaindia, 1995). When both words are in the same form (inflected or not), the compound should be written with a hyphen. The hyphen is also obligatory when the second word is inflected (*gorri-gorria*, *zuzen-zuzenean*). Componds such as *etxez etxe*, with the first word inflected are written without the hyphen. Sound-symbolic reduplications should be written with a hyphen (*nahas-mahas*). Carefully edited (and recent) texts tend to follow these rules.

3 Spell checkers

The two existing spell checkers for Basque are described in this section. We focus on their behavior with respect to the reduplication types we distinguish.

3.1 MSWord

Total reduplication Some cases are allowed, but it appears that there is a fixed list (e.g. *emeki-emeki* 'softly', *zuzen-zuzen* 'direct(ly)' or *estu-estu* 'narrow' are marked as incorrect, but *astiro-astiro* 'slowly' is not). When there is no hyphen, the suggestion in all cases is to eliminate the repeated word.

Second word inflected When there is no hyphen: correct (they are apparently treated simply as different words); with hyphen: again looks like a fixed list (*bakar-bakarrik* 'alone' and *zabal-zabala* 'wide' are incorrect, but *alper-alperrik* 'in vain' is admitted). Moreover, in some cases both spellings, with and without hyphen, are accepted, e.g. *bete betean* and *bete-betean* 'totally'). Normally the whole paradigm is marked as incorrect, e.g. in the case of *zuzen-zuzen* 'direct(ly)': *zuzen-zuzen-a*, *zuzen-zuzen-ean*, *zuzen-zuzen-etik*, *zuzen-zuzen-ak*, etc. (but not always: *oso-osoa* 'whole' is incorrect, but *oso-oso-rik* is not).

First word inflected When there is no hyphen: correct, with hyphen: incorrect.

Sound symbolic reduplication When there is no hyphen, second word is marked as incorrect; with hyphen it again seems like a fixed list (*erran-merran* 'gossip' is accepted, although *kokolo-mokolo* 'silly' or *esan-mesan* 'gossip' are not).

Lexical Flagged as incorrect.

Repetition of verbs Flagged as incorrect.

3.2 XuxenIV (for OpenOffice)

Total reduplication, syntactic and lexical reduplication In general, the spell checker does not treat repeated words as errors. Moreover, it does not deal with punctuation at all, and so all reduplications are correct whatever the spelling (as long as the words are in the lexicon)—*banan banan* and *banan-banan* 'one by one', *poz pozik* and *poz-pozik* 'happy', *etxez etxe* and *etxez-etxe* 'from one house to another'.

Sound symbolic reduplication It appears that a fixed list is used: e.g. *kokolo-mokolo* 'silly' is correct but *erran-merran* 'gossip' is not.

3.3 Evaluation

Xuxen does not treat repeated words as incorrect, which means that the problem is ignored and no difference is drawn between accidental repetition of words (user error) and instances of reduplication. MSWord, on the other hand, attempts to deal with the issue. It marks as incorrect all repetitions of identical phrases. However, some reduplications seem to be included in the spell checker lexicon as fixed phrases and in this case they are not treated as incorrect. The problem with this strategy is that it cannot fully deal with a productive process such as Basque reduplication. The only reduplication type that the spell checker handles well is the first word inflected type.

4 Duda-muda

This section presents our prototype, *Duda-muda.* It evaluates pairs of contiguous words and marks the pairs that fulfil at least one of the conditions. Possible outputs that the user is provided with are the following: correct, missing hyphen, hyphen not necessary or missing comma. It uses the following rules (in parentheses the type of reduplication the rule deals with and an example):

1. If the words are the same, there should be a hyphen (total reduplication, *gorri-gorri*).
2. If the second word contains the first word, there should be a hyphen (second word inflected, *gorri-gorria*).
3. If the second word differs from the first in that it has the prefix *m-*, there should be a hyphen (sound symbolic reduplication, *erran-merran*).
4. If the words differ in the first consonant (*m-* in the second word), there should be a hyphen (sound symbolic reduplication, *zehatz-mehatz*).
5. If the first word contains the second word, no hyphen is expected (first word inflected, *etxez etxe*).
6. If the two words are uninflected participles, there should be a comma (repeti-

tion of verbs, *ikusi, ikusi*; the database contains a list of verbs).

7. Special rules for some frequent cases of lexical reduplication (*bai, bai* 'yes, yes', *zoaz, zoaz* 'go, go').

4.1 Initial results

We tested our application with three texts: an extract from a novel by M. Ugarte (25,771 words), a novel by J. Satrustegi (23,741 words) and a collection of news articles from two Basque newspapers (*Berria* and *Gara*, 15,494 words). They contained 179 different cases of reduplication.

As the prototype does not have a list of admissible reduplications, it is able to handle not only frequent expressions, but also more innovative uses (e.g. *bonbardaketak bonbardaketa* 'despite the bombings', *herrikoi-herrikoia* 'very folk', *astinduaz-astinduaz* 'shaking'). Also, we do not impose any limitations on the word class the reduplicated words must belong to and this strategy allows infrequent cases to be dealt with, as it seems that at least morphological reduplication might be applicable to all parts of speech.

27 (15%) cases were classified erroneously. Most errors (20 out of 27) stem from the fact that sometimes two words that fulfil one of the conditions are not in fact the same lexeme, e.g. *ni nintzen* 'I was', *ezer ez* 'nothing' or *baina bai* 'but yes' and they are erroneously treated as reduplication. So far, the program cannot differentiate between inflectional endings and a part of a word—inflectional ending in the program is just that part of the word which the other word does not contain, e.g. in *gorri-gorriarekin* it is *-arekin*, which is correct, but in *bi bide* it is *-de*, which is in fact a part of the word. The problem would not exist if our prototype was a part of a general purpose spell checker that is able to recognize when two words correspond to the same lemma. A related problem is the case of expressions such as *laguna lagunarekin* 'the friend with the friend': the two words correspond to the same lemma, they look like a reduplication (similar to *gorri-gorriarekin*, for example), but they are not. In several cases the program suggests using a hyphen, but what in fact is needed, is a comma (e.g. *badator badator* 'he's coming, he's coming').

Another problem with our application is that all repetitions are treated as reduplications, i.e. when the user accidentally types two identical words, the program will suggest correction with a hyphen, rather than eliminating the repeated word. On the other hand, in the case of user error, since there is a reaction by the program when a repeated word is encountered, the user will realize that the repetition is a mistake.

5 Conclusions and future work

In this paper we propose a way of handling Basque reduplication in spell-checkers. We provide a classification of different types of reduplication and subsequently a set of rules that deal with each type. Tests show that our approach yields good results with most cases of reduplication. The application should be a module within a larger spell checking system which would ideally include a lexicon, a morphological module and, if possible, a syntax. Therefore, future work should be directed towards such an integration, more testing and completing the application with all the changes needed for improvement.

References

Agirre, E., I. Alegria, X. Arregi, X. Artola, A. D de Ilarraza, M. Maritxalar, K. Sarasola, and M. Urkia. 1992. XUXEN: a spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the third conference on Applied natural language processing*, pages 119–125.

Euskaltzaindia. 1995. Hitz elkartuen osaera eta idazkera. http://www.euskaltzaindia.net/dok/arauak/Araua_0025.pdf.

Hualde, J. I. 2003. Compounds. In J. I Hualde and J. O de Urbina, editors, *A Grammar of Basque.* Walter de Gruyter, Berlin-New York, pages 351–362.

Wiltshire, C. and A. Marantz. 2000. Reduplication. In G. Booij, C. Lehman, and J. Mugdan, editors, *Morphology: An international handbook on inflection and word-formation.* Walter de Gruyter, Berlin-New York, pages 557–567.