Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 266 – 277

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Building corpus-based frequency lemma lists

David Lindemann[a]*, Iñaki San Vicente[b]

[a]*UPV/EHU University of the Basque Country*
[b]*Elhuyar Foundation*

Abstract

This paper presents a simple methodology to create corpus-based frequency lemma lists, applied to the case of the Basque language. Since the first work on the matter in 1982, the amount of text written in Basque and language resources related to this language has grown exponentially. Based on state-of-the-art Basque corpora and current NLP technology, we develop a frequency lemma list for standard Basque. Our aim is twofold: On the one hand, to propose a primary Basque lemma list for a bilingual dictionary that is currently being worked on at UPV/EHU, and on the other, to contrast existing Basque dictionary lemma lists with frequency data, in order to evaluate the adequacy of our proposal and to compare lemma lists with each other.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Lexicography; Corpus Linguistics; Lemma Frequency; Basque Language

1. Introduction

Lexicography today resorts to corpora in order to find new dictionary entries; but how to decide whether a term is important enough to appear in a dictionary? Natural Language Processing techniques offer a chance to gather statistical information about words, such as information about the usage of candidate headwords, which helps us to gain evidence for a need to include it in a dictionary macrostructure. The evidence mostly referred to is frequency. For lexicographers, frequency data is important in three regards:

* Corresponding author. Tel.: +34-945-013-148; fax: +34-945-013-200.
E-mail address: david.lindemann@ehu.eus

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.445

1. The usage frequency of a lemma, which we can measure with corpus methods, is related to the look-up frequency of that lemma in dictionaries (De Schryver et al. 2010; Wolfer et al. 2014);
2. Frequency data is useful information for a lexicographer in the dictionary editing process;
3. Frequency data may be included in a dictionary microstructure (that is, the body of a dictionary entry), so that the dictionary user gets access to it. This is particularly useful for dictionary users who look up words in their L2. English learners, for instance, from the 1990s onwards find frequency information in dictionaries that have been designed for them (Kilgarriff 1997).

For this study, we have developed a frequency lemma list for Basque, following the model of state-of-the-art frequency lemma lists for German that are published under the title DeReWo. We have followed a double motivation:

1. to define a Basque lemma list for a first edition of the Basque→German part of EuDeLex, a bilingual dictionary currently being developed at UPV/EHU, and
2. to survey and prove criteria for including headword candidates in a dictionary lemma list, and to propose a methodology.

Having these goals in mind, we have assembled frequency lemma lists extracted from corpora, and examined the appropriateness of the evidence gained from these data sets. Furthermore, we have had the chance to compare the corpus-based lemma lists with the lemma lists of some Basque dictionaries. It is important to point out again that the aim of our work has not been to enrich an existing dictionary with more entries but to set up a the macrostructural content for a Basque dictionary from scratch.

This paper is organized as follows: In the following part, we give a short survey of the investigation on frequency lemma lists. In part 3, we present the language resources and the methodology used in this study. Part 4 is dedicated to the experiments we carried out and the results we have obtained, and part 5 offers reflections about these, some conclusions, and an outlook on future work on this issue.

2. The state of the art

2.1. German corpus-based frequency lemma lists

In order to provide resources for lexicography and other branches of linguistics, the Mannheim-based *Institut für Deutsche Sprache* publishes frequency word form and lemma lists under the title DeReWo, based on the DeReKo corpora (see Kupietz et al. 2010), which contain literary, scientific and press text from 1980 onwards. In 2014, these corpora counted 25 billion tokens.

Raw frequency data extracted from corpora to be valuable as headword candidate list in lexicography, some automatic, semi-automatic and handmade working steps are necessary, as frequency lists built entirely by automatic methods do not contain only accurate data. On the one hand, in order to assign a <lemma> or <lem-pos> (lemma and part of speech) pair to every word form on the list —in other words, to reach from word form frequency data to lemma frequency— the corpus has to be lemmatized and furnished with morphosyntactic tags, which is done by a linguistic tagger. On the other hand, for a lemma as headword candidate, a minimum frequency threshold has to be defined, that is, a minimal count of usage examples a lexicographer needs for defining the headword's semantic value or values (Sinclair 2005). For unigrams (single word lemmata), this threshold has been set to 20 occurrences (*ibid.*), but more factors have to be taken into account, such as the polysemy of a lemma and the number of homographs to it and their parts of speech.

In the case of DeReWo (see IDS 2009), the general method for building headword lists from frequency data can be summarized as follows:

1. The intersections of the frequency lemma list extracted from corpora with previously existing dictionary lemma lists are taken as accurate headword candidates;
2. The remaining list entries, those absent from previously published dictionaries, are classified by means of semi-automatic methods (that is, in groups) or by hand as accurate headword candidates or not.