Linking Basque Lexical Resources

Abstract

This poster presents preliminary considerations for a new project: A merged set of Basque (legacy) lexical resources or unified lexical database. At this preliminary stage, our main attention lies on the catalogue of data sources, on philological problems (e.g. regarding lemmatization), and on the design of the database. We propose a data model and a workflow for the inclusion of all kinds of Basque dictionaries and other resources such as the Basque WordNet and NLP lexicons. The data found in Basque dictionaries bare several problems, such as presence of dialectal and historical forms from before and after the creation of a Basque standard in 1968, and inconsistencies in lemmatization and the treatment of homography, homonymy, and polysemy. We present some solutions for these problems, as well as an XML schema for the lexical database and example datasets. The new merged resource may be used as Basque diachronic lexicographical corpus or serve as datasource for the creation of new lexicographical products. Using word sense identifiers found in Basque WordNet, the resource may also be linked to lexical resources of other languages.

Keywords: Historical Lexicography, Linking Lexical Resources, Basque Language

POSTER PROPOSAL – Language: English

1 Introduction

Two dates have a powerful meaning in Basque lexicography: On the one hand, the foundations for the creation of Standard Basque date from not earlier than 1968. On the other hand, from around 2005 Basque state-of-the-art electronic dictionaries and other resources have been published, and printed dictionaries are today fairly out of use and due to disappear. We propose to gather the data from Basque (legacy) lexicographical publications in an electronic database, with multiple goals in mind: (1) To preserve the lexicographical work done by numerous Basque scholars, (2) to make this work universally accessible, (3) to present all the data together, defining a common data model for this "dictionary of dictionaries", which shall allow the linking of objects stemming from legacy dictionaries from before 1968, i.e. the macrostructure of which does not obey Standard Basque lemmatization, through (a) Standard Basque lemma signs, (b) Standard Basque lempos-entities (i.e., lemmata with a unique part of speech value assigned), and (c), WordNet concepts (i.e. word sense identifiers found in Basque WordNet (EUSWN, *cf.* Table 1 below).

2 The Catalogue

In Table 1, we list Basque dictionaries and lexical resources of other types.[1] The tables' four columns represent the following:

1 The poster proposed here will contain sample pictures of paper dictionary content.

1. An abbreviation for each resource, typically consisting of a group of letters indicating title and/or authors;
2. The methods that according to preliminary assumptions apply for an inclusion of the data contained in the respective resource in the merged resource. The different methods are abbreviated as follows:
   a) TP Transcription and Parsing (i.e., representation in an XML structured format)[2], for manuscripts and older printed material
   b) OP Optical Character Recognition (OCR) and Parsing
   c) P Parsing only, if the data are already accessible in an electronic format
   d) X Transformations of the structural markup the data contains, for data already parsed or initially designed as electronic resource enriched with structural markup
3. An indication, whether the lemma list of the respective resource reflects nowaday's Standard Basque (YES) or it does not or does so partially (NO).
4. A short version of the bibliographical reference.

At the moment, data from about a dozen of these sources is already accessible to us in an electronic format and has been imported in a preliminary version of the merged resource.

Dictionary	**Method**	**Standard**	**Reference**
ADO1980	OP	NO	Kintana et al. (1980): *Hiztegia 80: euskara-espainiera, espainiera-euskara, vasco-español, español-vasco*
ADO1984	P	NO	Kintana et al. (1984): *Hiztegia bi mila: euskara-espainiera espainiera-euskara, vasco-español español-vasco*
ADO1986	P	NO	Uribarren et al. (1986): *Euskararako Hiztegia*
ADO1993	P	NO	Uribarren et al. (1993): *Europa hiztegia: eskola berrirakoa*
ADO1996	P	YES	Uribarren et al. (1996): *3000 Hiztegia*
ADO2009	X	YES	Adorez Taldea (2009): *Adorez 5000 Hiztegia*
ADO2013	X	YES	Bostak Bat Kultur Elkartea (2013): 5000 Hiztegia.
AIZ1883	OP	NO	Aizquibel (1883). *Diccionario basco-español titulado euskeratik erderara bi-urtzeco itztegia*
AUWH1992	P	NO	Aulestia & White (1992): *Euskara-Ingelesa Ingelesa-Euskara Hiztegia*. Donostia: Elkar
AZK1905	OP	NO	Azkue (1984 [1905-06]). *Diccionario vasco-español-francés*. Bilbo: Euskaltzain-dia [1905-06ko lehenengo argitalpenaren faksimilea]
BEME1916	OP	NO	Bera & Mendizabal (1916). *Diccionario castellano-euzkera / Bera'tar Eroman Mirena aba, buruñurduna. Euzkel-erdel iztegia / López Mendizabal'dar Ixaka*
BERA1909	OP	NO	Bera (1909). *Euzkel-iztegitxua / Bera'tar Erroman Aba*
CHA2002	X	NO	Charpentier (Charpentier 2002 [1823]): *Wörterverzeichnis, „in der zu Saint Etienne (Donostiÿ) Hauptdorf des Thales von Baigorrÿ bei Saint-Jean-Pied-De-Port üblichen Mundart“*

2 The authors have performed a series of tests for parsing the content of OEH dictionary using GROBID-dictionaries, a tool that relies on Conditional Random Fields, a Machine Learning approach, for parsing macrostructure and microstructural content, with promising results (cf. Blog Post [ANONYMIZED], 2017).

Dictionary	Method	Standard	Reference
EAH2007	P	YES	Martínez Rubio (2007): Euskara-Alemana Hiztegia.
EDBL	X	YES	Aldezabal et al. (2001) *Euskarazko Datu-Base Lexikala*
EEF2012	P	YES	Elhuyar Hiztegiak: Euskara-Frantsesa, 2. arg. (Pikabea et al. 2007) http://hiztegiak.elhuyar.org
EEG1996	P	YES	Elhuyar Hiztegiak: Euskara-Gaztelaina, 1. arg. (Azkarate et al. 1996)
EEG2000	P	YES	Elhuyar Hiztegiak: Euskara-Gaztelaina, 2. arg. (Azkarate et al. 2000)
EEG2006	T	YES	Elhuyar Hiztegiak: Euskara-Gaztelania, 3.arg. (Azkarate et al. 2006)
EEG2013	T	YES	Elhuyar Hiztegiak: Euskara-Gaztelaina, 4. arg. (Elhuyar Hizkuntza Zerbitzuak 2013) http://hiztegiak.elhuyar.org
EEN2007	T	YES	Elhuyar Hiztegiak: Euskara-Ingelesa, 2. arg. (Elhuyar Hizkuntza Zerbitzuak 2007) http://hiztegiak.elhuyar.org
EER1997	P	YES	Elhuyar Hiztegiak: Euskara-Errusiera, 1. arg. (Serrano et al. 1997)
EHA2012	X	YES	Euskaltzaindia (2012): *Euskaltzaindiaren hiztegia: Lesartk eta adibideak*
EM1987	OP	NO	Etxebarria, J.M. & Mujika, J.A. (1987). *Euskararen oinarrizko hiztegia: maiztasun eta prestasun azterketa*.
ETX2001	P	YES	Etxeberria (2001): *Ikaslearen Hiztegia*
EUNL1996	OP	YES	Jansen, W.H. (1996): Baskisch-Nederlands, Nederlands-Baskisch = Euskara-nederlandera, nederlandera-euskara
EUSWN	X	YES	*Euskal WordNet* (Pociello 2007; Pociello et al. 2011)
HAR1741	TP	NO	in: Harriet (1741): *Gramatica escuaraz eta francesez, composatua francez hitzcunça ikhasi nahi dutenen faboretan*.
HB2000	P	YES	Euskaltzaindia (2000): *Hiztegi Batua*
HB2010	X	YES	Euskaltzaindia (2010): *Hiztegi Batua*
HUM1817	TP	NO	Humboldt (1817): *Berichtigungen und Zusätze...*
IKAS1976	OP	NO	Montiano & Urquijo (1976): *Diccionario "IKAS" Euskera-Castellano Castellano-Euskara. Dialectos vizcaíno y guipuzcoano*
KÜH1999	OP	NO	Kühnel (1999): *Wörterbuch des Baskischen*
LAR1745	TP	NO	Larramendi (1745): *Diccionario Trilingüe*
LECL1826	OP	NO	in: Lécluse (1826): *Grammaire Basque*
LHA1926	OP	NO	Lhande (1926). *Dictionnaire basque-français et français-basque: dialectes Labourdin, Bas-Navarrais et Souletin*
LÖP1968	OP	NO	Löpelmann (1968): *Ethymologisches Wörterbuch der baskischen Sprache*
LUR1996	OP	YES	Arratibel, Atela & Navarro (1996): *LUR Eskolarako Hiztegi Entziklopedikoa*
MAH1840	TP	NO	Mahn (1840) [ms.]: Baskisches Wörterbuch I & II
MEND1962	OP	NO	López-Mendizabal (1962). *Diccionario vasco-español*, 4. arg.
MOD1977	OP	NO	Kintana (1977): *Euskal Hiztegi Modernoa*
MOD2000	P	YES	Barandiaran & Etxeberria (2000): *Euskal Hiztegi Modernoa*, 2. arg. berritua
MORR1998	X	YES	Morris (1998): Morris Student Plus. Euskara-Ingelesa, English-Basque
MUG1981	P	NO	Múgica Berrondo (1981): *Diccionario vasco-castellano*
OEH	X	NO	Euskaltzaindia (Mitxelena & Sarasola 1988): *Orotariko Euskal Hiztegia*
POUV1665	TP	NO	Pouvreau (1665) [ms.]: *Euskara-frantses(-latin-gaztelaniazko) hiztegia*.
RET1976	OP	NO	Sota, Lafitte & Akesolo (1976): *Diccionario Retana de autoridades de la lengua vasca: con cientos de miles de nuevas voces y acepciones, antiguas y modernas*
SEH1996	P	YES	Sarasola (1996): *Euskal Hiztegia*

Dictionary	Method	Standard	Reference
SEH2007	P	YES	Sarasola (2007): *Euskal Hiztegia*
SEIH1999	P	YES	Sarasola (1999): *Euskara Ikaslearen Hiztegia*
SHLH1994	X	YES	Sarasola (1994): *Hauta-lanerako Euskal Hiztegia*. http://www.euskara.euskadi.net/r59-sarasola/eu/sarasola/sarasola.apl
SMH1982	X	NO	Sarasola (1982). *Gaurko euskara idatziaren maiztasun-hiztegia: 1977ko corpus batean oinarritua*
UZEI1987	X	YES	Euskalterm (UZEI 1987): http://www.euskara.euskadi.net/r59-euskalte/eu/q91EusTermWar/kontsultaJSP/q91aAction.do
UZEI2004	X	YES	UZEI (2004): *Maiztasun Hiztegia*
UZEI2006	X	YES	UZEI (2006): *Atzekoz aurrera. Hitz-bukaeren hiztegia*
ZEH2005	X	YES	Sarasola (2005): *Zehazki: gaztelania-euskara hiztegia.* http://www.ehu.eus/ehg/cgi/zehazki

Table 1: Basque lexical resources and methods for inclusion in a merged resource

3 Data Modelling

As further development of the proposal presented in ([ANONYMIZED] et al., 2016], we propose to build a single XML document[3] with a DTD following the TEI-P5 guidelines. On the first hierarchy level, a list of Standard Basque lemma signs is set as macrostructure for the merged resource (TEI-P5 element <super-entry>). The entry-headwords from the different lexical resources to become part of the merged resource will be inserted as child elements hereof (TEI-P5 element <entry>). Below each of these, the respective microstructural content is to be inserted, with a microstructural markup as rich as possible, from case to case. Whenever possible, i.e. if the data of a particular source is accessible in or can be transformed to proper XML, we will establish a syntactical entity which carries the part-of-speech (POS) information and, at the same time, serves for disambiguating homographs and homonyms (i.e. homographs with the same POS), so that the other information contained in the dictionary articles (semantic information etc.) is stored inside an element that is child to that syntactical entity. This modelling is particularly interesting for Basque, since 5% of lemma signs have more than one part of speech assigned.

We are sure the Galway workshop will be a very good opportunity to discuss this approach of data modelling, and the technologies to be employed for retro-digitizing Basque legacy dictionaries, and to learn more about technologies available through or to be developed by the ELEXIS project.

3 The poster proposed here will contain figures representing the structure of that XML document.