BASQUE
E-LEXICOGRAPHIC RESOURCES: LINGUISTIC BASIS, DEVELOPMENT, AND FUTURE PERSPECTIVES

Izaskun Aldezaabal, Xabier Artoloa, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios, Gorka Labaka, German Rigau, Ruben Urizar
Ixa Group, University of the Basque Country (UPV/EHU)

**ACKNOWLEDGEMENTS**

TUNER: Automatic domain adaptation for semantic processing (TIN2015-65308-C5-1-R), funded by the Spanish Ministry of Economy and Lexikoaren Behatokia IX, funded by Euskaltzaindia (Academy of the Basque Language)

**FUTURE PERSPECTIVES**

Add the terms that are used frequently in most of the science areas in both resources. Moreover in BWN:

- WNTERM: terms from the Science and Technology dictionary
- Terms found in the logbooks (nautical)
- Difficult to decide its place in BWN because of too specialised words
- No equivalent in PWN -> use CILI

- Semi-automatic approaches to update the resources:
  - Similarity measures to compare place names in EDBL-LBDBL with foreign and Basque spellings of the same location: *Philadelphia vs Filadelfia*
  - Black-box techniques by cross-checking different ontologies with WordNet to detect knowledge discrepancies

**LINGUISTIC ISSUES: LEXICALISATION**

Deciding Basque Word Forms

Do not add regular forms that can be derived from the lemma and that can also be analysed from simpler constituents:

- Lexical suffixes relating ordinals: *ba 'two' -> bigaren 'second'
- Intensity markers: *hau 'this one' -> hauze 'just this one'
- Possessive pronouns: *ni 'I' -> nire 'mine'
- Nouns used as postpositions or complementisers: *aurre 'front' -> aurrean 'in front of'
- Words with many spatio-temporal case markers: *meza 'mass' -> mezan, mezetan 'in mass'
- Words with modal case markers: *huts 'cold' -> hutsze 'be/feel cold'
- Verbal nouns: *egun 'do' -> gituze 'doing'
- Causative verbs: *egun 'do' -> egunazze 'make someone do'

Add forms specialised meaning: e.g. *endirizte 'centering', a verbal noun, but a term in football (providing specialised vocabulary, but not specifically coded)*

Deciding Concepts

Conceptual level imbalances vs expression level imbalances + Basque word form problems:

1. Merging PWN synsets: [actor, histrion, player, thespian, role_player] and [actress] -> [aktore, antzazle, komediante, antzezlar]
2. Splitting PWN synsets: [terrorist_organization, terrorist_group, foreign_terrorist_organization]
3. Adding Basque synsets: *enbata, Basque for sudden rough weather in the Bay of Biscay and in the Cantabrian Sea*

Procedure for multi-words not found in Basque dictionaries:

1. Create manually variant proposals based on translations of each unit
2. Look for them in the corpora
3. If found -> add with a special label (ixalex) e.g. *animalia-birus [animal, virus]; if not -> label the synset as non-lexicalised (nonlex) e.g. [cranio metric_point]*

**UPDATE PROCESS**

EDBL-LBDBL

- Source: Euskaltzaindiaren Hiztegia
- Detect the changes by checking automatically dictionary versions: new entries and subentries, changes in the standardisation mark and/or level, and deleted entries
- Populating procedure:
  - Entry with PoS: 1) heuristics to create database information (lemma, two-level form...) 2) manual revision
  - Entries without PoS: manually

BWN

- Source: PWN, Basque dictionaries, terminological databases, corpora, and Basque Wikipedia
- Upper concepts updated in version 3.0: the Base Level Concepts, the general concepts (genlex) and the eponyms
- Methodology defined by Pociello (2008): manual translation, but using updated referential resources and also checking the sense in Euskaltzaindiaren Hiztegia

**TWO E-LEXICOGRAPHIC RESOURCES AT IXA GROUP**

Standardisation of Basque language began officially 50 years ago -> challenging development and maintenance

**Monolingual lexical database with morphological information and normative information**

- Creation: 1992; massive population 2010; regular updates (every new version of Euskaltzaindiaren Hiztegia)
- Entry types: standard dictionary entries, non-standard variants linked to their standard equivalents, finite and irregular verb forms, dependent morphemes, compounds, multi-word entries, abbreviations, etc.
- Total entries: 135,062; dictionary entries: 113,682
- Main applications: the spell checker Xuxen and the automatic processing of Basque texts
- Non-commercial license, graphical user interface

**No equivalent in PWN**

- • Causative verbs:
- • Verbal nouns:
- • Difficult to decide its place in BWN because of too specialised words
- • Terms found in the logbooks (nautical)
- • WNTERM: terms from the Science and Technology dictionary
- • Words with modal case markers:
- • Words with many spatio-temporal case markers:
- • Lexical suffixes relating ordinals:
- • Nouns used as postpositions or complementisers:
- • Possessive pronouns:
- • Intensity markers:
- • Words with many spatio-temporal case markers:
- • Words with modal case markers:
- • Verbal nouns:
- • Causative verbs:
- • Words with many spatio-temporal case markers:
- • Words with modal case markers:
- • Verbal nouns:
- • Causative verbs:
- • Words with many spatio-temporal case markers:
- • Words with modal case markers:
- • Verbal nouns:
- • Causative verbs: