Ixa Group Linguistic Resources

Itziar Gonzalez-Dios
Taller ReTeLe
Salamanca, 13/09/2016

Corpora

- Euskararen Prozesamendurako Erreferentzia Corpusa (EPEC) - Reference Corpus for the Processing of Basque
- Euskal RST Treebank
- Zientzia eta Teknologia corpusa (ZT) - Science and Technology corpus
- QTLeap Multilingual Corpus
- Basque English ParDeepBank

Corpus: EPEC (Reference Corpus for the Processing of Basque)

- Linguality type: written
- Language: Basque
- Text format: xml
- Size
  - Number of sentences: 25.000
  - Number of words: 300.000
- Character encoding: UTF-8

Corpus: EPEC (Reference Corpus for the Processing of Basque)

- Classification
  - –Text type: journalistic, small part literary
  - –Register: neutral
- Annotation: segmentation, constituents, shallow syntax, dependencies, EuSemcor, Rolsem and coreference
- Creation
  - –Creation mode: semi-automatic, manually
- Version: None
- Documentation: Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. *Corpus Linguistics Around the World. Book series: Language and Computers. Vol 56 (pag 1- 15). ISBN 90-420-1836-4 Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands.*

Distribution

- Availability: according to the annotation level/ partial corpus
- License
  - License:
    - Dep: CC BY-NC-SA 3.0
    - Eusemcor: contact us
    - Korref: CC BY 4.0
  - Restrictions: -
  - Download location:
    - Dep http://ixa.si.ehu.es/Ixa/Produktuak/deskargaform?ida=1306407157,
    - EuSemcor *http://ixa.si.ehu.es/Ixa/Produktuak/deskargaform?ida=1306407157*,
    - Korref http://ixa2.si.ehu.eus/epec-koref/epec-koref_v1.0.tgz
  - Distribution Access/Medium: digital-downloadable

Contact Person

DEP : María Jesús Aranzabe maxux.aranzabe@ehu.eus
EuSemcor: Eneko Agirre e.agirre@ehu.eus
Korref : Ander Soraluze ander.soraluze@ehu.eus

Resource Creation

- Resource creator: Itziar Aduriz, Eneko Agirre, Izaskun Aldezabal, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Klara Ceberio, Arantza Díaz de Ilarraza, Ainara Estarrona, Jone Etxeberria, Kike Fernandez, Uxoa Iñurrieta, Mikel Iruskieta, Elixabete Izagirre, Karmele Mendizabal, Eli Pociello, Ander Soraluze, Larraitz Uria
- Funding project: CESS-ECS (MEC) 2005; HIZKING21 (Basque Gov.)

Corpus: RST Basque Treebank

- Linguality type: written
- Language: Basque
- Text format: txt, rs3
- Size
  - Number of sentences: -
  - Number of words: 24.057
  - Elementary discourse units: 2.509
- Character encoding: UTF-8

Corpus:: RST Basque Treebank

- Classification
  - Text type: medical, terminological and scientific
  - Register: neutral
- Annotation: discourse relations following RST
- Creation
  - Creation mode: manual
- Version: Basque RST TB is a corpus in evolution
- Documentation: Iruskieta M., Aranzabe M. J., Diaz de Ilarraza A., Gonzalez-Dios I., Lersundi I., Lopez de Lacalle O. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations. *4th Workshop RST and Discourse Studies, 40-49, Sociedad Brasileira de Computacao, Fortaleza, CE, Brasil. October 20-24*

Distribution

- Availability: interface http://ixa2.si.ehu.es/diskurtsoa/en/
- License
  - License: thinking
  - Restrictions: -
  - Download location: http://ixa2.si.ehu.es/diskurtsoa/fitxategiak.php
  - Distribution Access/Medium: digital-downloadable

Contact Person: Mikel Iruskieta mikel.iruskieta@ehu.eus

Resource Creation

- Resource creator: María Jesús Aranzabe, Arantza Díaz de Ilarraza, Kike Fernandez, Itziar Gonzalez-Dios, Mikel Iruskieta, Mikel Lersundi, Oier Lopez de Lacalle
- Funding project: Ixa Group, Research group of type A (Basque Gov.)

Corpus: ZT corpusa

- Linguality type: written
- Language: Basque
- Text format: XML, TEI-P4
- Size
  - Number of sentences: -
  - Number of words: 1.9 million balanced, manually revised: 6 million
- Character encoding: UTF-8

Corpus: ZT corpusa

- Classification
  - –Text type: scientific and technologycal
  - –Register: neutral
- Annotation: morpho-syntactic
- Creation
  - –Creation mode: semi-automatic

Version: none

- Documentation: N. Areta, A. Gurrutxaga, I. Leturia, Z. Polin, R. Saiz, I. Alegria, X. Artola, A. Diaz de Ilarraza, N. Ezeiza, A. Sologaistoa, A. Soroa, A. Valverde. 2006. Structure, Annotation and Tools in the Basque ZT Corpus. *LREC-2006*. ISBN 2-9517408-2-4

Distribution

- Availability: interface http://www.ztcorpusa.eus/cgi-bin/kontsulta.py

Contact Person: Xabier Artola xabier.artola@ehu.eus

Resource Creation

- Resource creator: Izaskun Aldezabal, Iñaki Alegria, Arrate Andres, Xabier Artola, Arantza Díaz de Ilarraza, Ainara Estarrona, Jone Etxeberria, Nerea Ezeiza, Kike Fernandez, Mikel Iruskieta, Izaskun Izagirre, Mikel Lersundi, Aitor Sologaistoa, Andoni Valverde, Nerea Areta, Antton Gurrutxaga, Igor Leturia
- Funding project: Hizking21 (Basque Gov.; Gipuzcoan Foral Gov.)

Corpus: QTLeap Corpus

- Linguality type: written
- Language: multilingual, parallel (Spanish, Czech, Dutch, English, Portuguese, Bulgarian, Basque, German)
- Text format: txt
- Size
  - Number of sentences: 9.959 (4.000 questions & 4.000 aswers)
  - Number of words: 139.411 (EN)
- Character encoding: UTF-8

Corpus: QTLeap Corpus

- Classification
  - –Text type: questions and answers pairs, computer and IT troubleshooting domain
  - –Register: naturally occurring utterances
- Annotation: none
- Creation
  - –Creation mode: manual translation
- Version: -
- Documentation: http://qtleap.eu/wp-content/uploads/2015/05/QTLEAP-2015-D2.51.pdf

Distribution

- Availability: Available- restricted use
- License
  - –License: CC–BY-NC-SA
  - –Restrictions: -
  - –Download location: http://metashare.metanet4u.eu/repository/browse/qtleap-corpus/0176c39ae9cd11e4a2aa782bcb074135ba7d767f645a48dca1d50ee3c9504253/
  - –Distribution Access/Medium: digital-downloadable

Contact Person: Rosa Del Gaudio

Resource Creation

- Resource creator: Rosa Del Gaudio
- Funding project: QTLeap European Project

Corpus: Basque English ParDeepBank

- Linguality type: written
- Language: Basque-English (parallel)
- Text format: txt, CONLL format
- Size
  - Number of sentences: 805
  - Number of words: 16.002 (8.388 EN; 7.614 EU)
- Character encoding: UTF-8

Corpus: Basque English ParDeepBank

- Classification
  - Text type: journalistic
  - Register: neutral
- Annotation: lemmatization, morphological analysis, dependency parsing trees
- Creation
  - Creation mode: semi-automatic
- Version: 1.0
- Documentation: http://qtleap.eu/wp-content/uploads/2016/01/QTLEAP-2015-D4.10.pdf

Distribution

- Availability: Available-restricted use
- License
  - License: CC-BY
  - Restrictions: -
  - Download location: http://metashare.metanet4u.eu/repository/browse/basque-english-pardeepbank/1c665572104111e5a2aa782bcb074135247c22693c2e4320891b8feca50751e7/
  - Distribution Access/Medium: digital-downloadable

Contact Person: Eneko Agirre e.agirre@ehu.eus

Resource Creation

- Resource creator: Ainara Estarrona, Arantza Otegi, Larraitz Uria
- Funding project: QTLeap European Project

Lexical Conceptual Resources

- Euskararen Datubase Lexikala (EDBL) - Lexical database of Basque
- Basque Verb Index (BVI)
- Multilingual Central Repository (MCR)
- Basque WordNet

Lexical Conceptual Resource: EDBL

- Resource type: lexical database
- Linguality type: written
- Languages: Basque
- Text format: txt, xml
- Size
    - Number of entries: 132.350
- Character encoding: UTF-8

Lexical Conceptual Resource: EDBL

- Domains: general
- Encoding
  - Encoding level: lexical, morphological
  - Linguistic information: lemma, PoS, syntactic functions
- Creation mode: manually, semi-automatic
- Version: 5
- Documentation: Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi M. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop on linguistic databases.* Philadelphia (USA).

Distribution

- Availability: interface http://ixa2.si.ehu.es/edbl
- License
  - License: ELRA
  - Restrictions: -
  - Download location: http://catalog.elra.info/product_info.php?products_id=804
  - Distribution Access/Medium: digital-downloadable

Contact Person: Gorka Labaka gorka.labaka@ehu.eus

Resource Creation

- Resource creator: Ixa Taldea
- Funding project: Ixa Group, Research group of type A (Basque Gov.)

Lexical Conceptual Resource: Basque Verb Index (BVI)

- Resource type: corpus-based lexicon
- Linguality type: written
- Languages: Basque
- Text format: database
- Size
  - Number of entries: 270 (verbs with 30 occurences at least, covering the % 85 of the EPEC corpus)
- Character encoding: -

Lexical Conceptual Resource : Basque Verb Index (BVI)

- Domains: general
- Encoding
  - Encoding level: semantics
  - Linguistic information: Basque verb and its PropBank equivalent, semantic roles in VN-PB and EADB (database for Basque Verbs); linked to PropBank, Verbnet, WordNet, Levin's classification and FrameNet
- Creation mode: semi-automatic
- Version: none
- Documentation: Estarrona A., Aldezabal I., Díaz de Ilarraza A. and Aranzabe M.J. 2015. Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/Verbnet model. *Edward Vanhoutte (ed.) Digital Scholarship in the Humanities, Volume 30, Number 2, 1-23. Oxford University Press (Online ISSN 2055-768X - Print ISSN 2055-7671) doi: 10.1093/llc/fqv001*

Distribution

- Availability: interface http://ixa2.si.ehu.es/e-rolda/index.php?lang=en
- License
  - License: thinking
  - Restrictions: only online
  - Download location: -
  - Distribution Access/Medium: digital

Contact Person: Ainara Estarrona ainara.estarrona@ehu.eus

Resource Creation

- Resource creator: Ainara Estarrona
- Funding project: Ixa Group, Research group of type A (Basque Gov.)

Lexical Conceptual Resource: Multilingual Central Repository (MCR)

- Resource type: multilingual lexical database with wordnets
- Linguality type: written
- Languages: English, Spanish, Catalan, Basque, Galician and Portuguese
- Text format: tsv
- Size
  - Number of entries: more than one million semantic relations
- Character encoding: UTF-8

Lexical Conceptual Resource: Multilingual Central Repository (MCR)

- Domains: several
- Encoding
  - Encoding level: semantics
  - Linguistic information: WN, semantically tagged glosses, WN domains, Base Concepts, Top Ontology, AdimenSUMO ontology
- Creation mode: automatic mappings
- Version: 3.0
- Documentation: Gonzalez-Agirre A., Laparra E. and Rigau G. Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In Proceedings of the Sixth International Global WordNet Conference (GWC'12). Matsue, Japan. January, 2012.

Distribution

- Availability: interface http://adimen.si.ehu.es/web/MCR
- License
  - License: the English WordNet synset and relation data distributed under the original WordNet license. All other data CC BY 3.0.
  - Restrictions: -
  - Download location: http://adimen.si.ehu.es/web/
  - Distribution Access/Medium: digital-downloadable

Contact Person: German Rigau german.rigau@ehu.eus

Resource Creation

- Resource creator: Rodrigo Agerri, Aitor Gonzalez-Agirre and German Rigau
- Funding project: MEANING European project; KNOW, KNOW2 and Skater Spanish gov. projects

Lexical Conceptual Resource: Basque WordNet

- Resource type: lexical-semantic database
- Linguality type: written
- Languages: Basque
- Text format: tsv
- Size
  - Number of entries: 93.353 word senses; 59.948 words
- Character encoding: UTF-8

Lexical Conceptual Resource: Basque WordNet

- Domains: general
- Encoding
  - Encoding level: semantics
  - Linguistic information: synsets
- Creation mode: semi-automatic, manually reviewed
- Version 3.0
- Documentation: Pociello E., Agirre E. and Aldezabal I. 2011. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation, 45 (2).* Springer.

Distribution

- Availability: interface http://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl
- License
  - License CC BY 3.0
  - Restrictions: -
  - Download location: http://adimen.si.ehu.es/web/files/mcr30/mcr30.zip
  - Distribution Access/Medium: digital-downloadable

Contact Person: Eneko Agirre e.agirre@ehu.eus

Resource Creation

- Resource creator: Eneko Agirre, Izaskun Aldezabal, Olatz Ansa, Eli Pociello and German Rigau
- Funding project: Ixa Group, Research group of type A (Basque Gov.)

Ixa Group Linguistic Resources

Itziar Gonzalez-Dios
Taller ReTeLe
Salamanca, 13/09/2016