# CORPUS HISTORIKOEN PROZESAKETA
## Jardunaldi irekia, Ekainaren 11n

Corpus historikoak biltzea, etiketatzea, aztertzea eta kontsultagarri jartzea nahitaezkoa da hizkuntzaren eta kulturaren eboluzioa modu kuantitatiboan ikertu ahal izatea. Hizkuntzalaritza, historia eta teknologia arloen arteko lankidetza beharrezkoa da aipatutako prozesuak arrakastatsuak izango badira.

Nazioartean hainbat proiektu ari dira egiten ildo horretan eta jardunaldi honetan esperientzia horietako batzuk azalduko dira. Euskal Herrian ere hainbat proiektu daude martxan baina modu atomizatuan.

**Noiz**: 2018ko ekainaren 11, goizeko 11.00etan (Ada Lovelace aretoa)

**Non**: EHUko Informatika Fakultatea, Manuel Lardizabal 1, 20018 Donostia (mapa)

**Hizkuntza**: ingelesa

**Programa**:

- **11.00-11.30**: Ricardo Etxepare: *BIM project, Basque in the making (Sintaktikoki Etiketatutako Euskarazko Corpus Historikoa)*
- **11.30-12.15**: Martin Reynaert: *Text-Induced Corpus Clean-up: current state-of-the-art*
- **12.15-13.00**: Eckhard Bick: *Automatic Grammatical Annotation of Historical Brazilian Portuguese*

**Babesleak**: UPPA-UPV/EHU. Clarin.

# PROCESSING OF HISTORICAL CORPORA
## Open day. June 11th.

The collection, tagging, analysis and recovery of historical corpora are basic tasks in the quantitative research on linguistic and cultural evolution. Collaboration between the areas of linguistics, history and technology is necessary for the success of these processes.

Several international projects are being carried out in this field and some of these experiences will be presented at this workshop. In the Basque Country there are also projects in progress but in an atomised manner.

**Date**: June 11th. 11.00 a.m. (Ada Lovelace hall)

**Place**: Informatics Faculty UPV/EHU. Manuel Lardizabal 1, 20018 Donostia (map)

**Language**: English

**Program**:

- **11.00-11.30**: Ricardo Etxepare: *BIM project, Basque in the making (Sintaktikoki Etiketatutako Euskarazko Corpus Historikoa)*
- **11.30-12.15**: Martin Reynaert: *Text-Induced Corpus Clean-up: current state-of-the-art*
- **12.15-13.00**: Eckhard Bick: *Automatic Grammatical Annotation of Historical Brazilian Portuguese*

**Sponsor**: UPPA-UPV/EHU

**PROCESADO DE CORPUS HISTÓRICOS**
**Jornada abierta. 11 de Junio.**

La recopilación, etiquetado, análisis y consulta de corpus históricos son tareas fundamentales en la investigación cuantitativa de la evolución lingüística y cultural. La colaboración entre las áreas de lingüística, historia y tecnología es necesaria para el éxito de los procesos mencionados.

Diversos proyectos internacionales se están llevando a cabo en este ámbito y en esta jornada se expondrán algunas de estas experiencias. En Euskal Herria también hay proyectos en marcha pero de forma atómizada.

**Fecha**: 11 de junio de 2018, 11.00. (Sala Ada Lovelace)

**Lugar**: Facultad de Informática UPV/EHU. Manuel Lardizabal 1, 20018 Donostia (mapa)

**Idioma**: inglés

**Programa**:

- **11.00-11.30**: Ricardo Etxepare: *BIM project, Basque in the making (Sintaktikoki Etiketatutako Euskarazko Corpus Historikoa)*
- **11.30-12.15**: Martin Reynaert: *Text-Induced Corpus Clean-up: current state-of-the-art*
- **12.15-13.00**: Eckhard Bick: *Automatic Grammatical Annotation of Historical Brazilian Portuguese*

**Patrocinadores**: UPPA-UPV/EHU

**Text-Induced Corpus Clean-up: current state-of-the-art**

All legacy text corpus building and exploitation efforts around the world face the same massive hurdle: the text quality deterioriation caused by the digitization process, the noisy channel of Optical Character Recognition or OCR, which most texts necessarily need to pass.
We have gradually been further developing our OCR post-correction system 'Text-Induced Corpus Clean-up' or TICCL on the basis of the major evaluations we described in Reynaert (2016).
We will give an overview of TICCL's various modules and how these work and interact. We will give an introduction to TICCL's main lexical variant retrieval algorithm,. We will also explain how and why it should be perfectly feasible to quickly and likely successfully adapt TICCL to a new language, even a venerable pre-Indo-European language such as Basque.
The work dictated by TICCL's previous major evaluations is now nearing completion and we expect to be able to present first new evaluation results based on the major enhancements to TICCL effected, i.e. language detection, word ngram correction and chaining of correction candidates. We can now handle word splits and run-ons, have a better grip on short word errors and are able to retrieve and correct errors beyond Levenshtein distance 2 with great precision.
With a look towards the future we will sketch our plans for the next year in which we are set to build TICCLAT, the 'Text-Induced Corpus Correction and Lexical Assessment Tool', in collaboration with the Dutch eScience Center (1), drawing on the rich resources now available in the 20 billion word diachronic corpus of Dutch we helped to build in the Nederlab project (Brugman, 2016).

**Martin Reynaert**

Hennie Brugman, Martin Reynaert, et al. (2016) Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In Nicoletta Calzolari et al., editor, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016), pages 1277–1281, Portorož, Slovenia. ELRA. ISBN 978-2-9517408-9-1. URL http: //www.lrec-conf.org/proceedings/lrec2016/pdf/471_Paper.pdf.

Martin Reynaert. (2016) OCR post-correction evaluation of Early Dutch Books Online – revisited. In Nicoletta Calzolari et al., editor, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016), pages 967–974, Portorož, Slovenia. ELRA. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/596_Paper.pdf.

**Automatic Grammatical Annotation of Historical Brazilian Portuguese**

The talk discusses methods and challenges involved in the grammatical annotation of a corpus of historical Brazilian Portuguese representing the administrative area of São Paulo. Morphological anaysis and disambiguation as well as syntactic parsing were carried out with program modules adapted from a modern Portuguese parser, PALAVRAS, and in order to make this work, a complex, orthographical "translation" between historical and modern Portuguese had to be performed, involving systematic orthographical filtering, lexicon additions and morphological heuristics - a method shown to ultimately provide correct PoS tagging in 95-98%, and acceptable syntactic function tags in 91-94%, depending on text type. I will discuss the architecture of the system as well as specific problems involved in the historical-to-modern conversion of word forms, and present a qualitative comparison across text types, with statistical results from the annotated corpus.

**Eckhard Bick** is a computational linguist and project leader for the VISL lab at the University of Southern Denmark, where he works as a language technology researcher at the Department of Language and Communication (ISK). Over the years he has designed and developed grammars, corpora, lexical resources and applicational tools for a large number of languages, including most of the Romance and Germanic languages. Eckhard Bick is a leading expert in the field of Constraint Grammar, with a current focus on semantic annotation and machine translation. Eckhard Bick has published extensively on various aspects of computational linguistics and participated in a large

number of international research projects.