

# Koldo Gojenola

Koldo Gojenola (born 2/09/1965. ORCID num: 0000-0002-2116-6611) has a degree in Computer Science (1988) and received his doctorate in Computer Science in 2000. He is an associate professor. He credits 35 years of teaching experience until 2023 (6 five-year periods).

He has published more than 100 articles in international magazines and conferences in the area of Natural Language Processing (NLP). The main research objective has been the development of applications that help to process texts robustly, efficiently and with high precision and coverage, using linguistic and statistical methods, supervised and unsupervised.

Currently he works on several research areas, including morphological, syntactic and semantic processing of Basque, Spanish and English, and on the application of linguistic and statistical techniques to the processing of medical texts. The contributions include 33 publications in journals according to the Journal Citation Report (JCR), and 4 articles in high-level conferences (CORE A and A+).

He is a regular reviewer for prestigious international magazines such as LRE, and JNLE. He is a regular reviewer of CORE A and A\* conferences such as ACL, NAACL, and EMNLP. He has been the IP in projects funded by the Basque and the Spanish Government. He has also participated as a researcher in European projects (QTLeap, LoCloud, Antidote). In addition, he has participated in several national projects as Principal Investigator (DOTT-HEALTH, PROSA-MED, EXTRECM).

He has been organizer of several research conferences, as the president of the Annual Conference of the Spanish Association for Natural Language Processing (SEPLN) in 2009, and president of the local organizing committee of the International Conference on Parsing Technologies (IWPT) in 2015.

Teaching and research proficiency in Spanish, English and Basque.

## **General indicators of quality of scientific production (see instructions)**

Three positive research evaluations (sexenio) from the Spanish Ministry of Education.

Directed Phd theses: 4

Publications in Thomson Reuters Web of Science (total): 33

Publications in Thomson Reuters Web of Science (Q1): 5

Total cites (Google Scholar): 1970

h index (Google Scholar): 23

## **C.3. Contracts, technological or transfer merits**

Contract: Itzulbide: Desarrollo e implementación de un sistema de traducción automática neuronal para traducir informes clínicos entre el euskera y el castellano (G/211/20/1/0199/o001/0000/012019). Servicio Vasco de Salud, Osakidetza. Maite Oronoz Anchordoqui. 05/06/2019-05/06/2021. The Basque Health System (Osakidetza) is the contractor and it consists of the development of an automatic translator for the translation of real medical reports from Basque to Spanish. It has a duration of 24 months, a turnover of 191,469.58 € and the research team consists of 9 PhD researchers.

Registry of Intellectual property: In 2019, the software "IxaMedTagger: Automatic system for the automatic detection of entities in the medical domain in Spanish" was registered as a result of the research work developed in the field of information extraction with medical documents.

## **Selected publications (last 5 years)**

Authors: Iker Goenaga, Edgar Andrés, Koldo Gojenola, Aitziber Atutxa

Title: Advances in Monolingual and Crosslingual Automatic Disability Annotation in Spanish

Journal: BMC Bioinformatics volume 24, Article number: 265, 2023

**Quality indicator: JCR Q2**

Authors: Iker de la Iglesia, María Vivó, Paula Chocrón, Gabriel de Maeztu, Koldo Gojenola, Aitziber Atutxa

Title: An Open Source Corpus and Automatic Tool for Section Identification in Spanish Health Records

Journal: Journal of Biomedical Informatics, 2023

**Quality indicator: JCR Q2**

Authors: Iakes Goenaga, Xabier Lahuerta, Aitziber Atutxa, Koldo Gojenola

Title: A Section Identification Tool: towards HL7 CDA/CCR Standardization in Spanish Discharge Summaries

Journal: Journal of Biomedical Informatics, 2021

**Quality indicator: JCR Q1**

Authors: A. Atutxa, A. Díaz de Ilarraza, K. Gojenola, M. Oronoz, O. Perez de Viñaspre

Title: Interpretable Deep Learning to Map Diagnostic Texts to ICD10 Codes

Journal: International Journal of Medical Informatics, 2019

**Quality indicator: JCR Q2**

Authors: Jorge Pérez, Alicia Pérez, Arantza Casillas, Koldo Gojenola

Title: Cardiology record multi-label classification using Latent Dirichlet Allocation

Journal: Computer Methods and Programs in Biomedicine, 2018.

**Quality indicator: JCR Q1**

Authors: Alicia Pérez; Aitziber Atutxa; Arantza Casillas; Koldo Gojenola; Alvaro Sellart

Title: Inferred joint multigram models for medical term normalization according to ICD

Journal: International Journal of Medical Informatics, Volume 110, February 2018, Pages 111–117.

**Quality indicator: JCR Q2**

Authors: Alicia Pérez; Rebecka Weegar; Arantza Casillas; Koldo Gojenola; Maite Oronoz; Hercules Dalianis

Title: Semi-supervised medical entity recognition: a study on Spanish and Swedish clinical corpora

Journal: Journal of Biomedical Informatics, Volume 71, 2017, Pages 16–30,

**Quality indicator: JCR Q2**

### **Selected projects (last 5 years)**

Title: DOTT-HEALTH: Advanced semantic textual processing for the detection of diagnostic codes, procedures, concepts and their relationships in health records

Institution: Ministerio de Economía y Competitividad, Spain (TIN2016-77820-C3-1-R)

Duration: 06-2020 - 06-2023

Main researcher: Dr. Koldo Gojenola and Arantza Casillas

Title: PROSA-MED: Advanced semantic textual processing for the detection of diagnostic codes, procedures, concepts and their relationships in health records

Institution: Ministerio de Economía y Competitividad, Spain (TIN2016-77820-C3-1-R)

Duration: 01-2017 - 12-2019

Main researcher: Dr. Koldo Gojenola and Dra. Arantza Diaz de Ilarraza

Title: EXTRECM: EXTracción de RELaciones entre Conceptos Médicos en fuentes de información heterogéneas

Institution: Ministerio de Economía y Competitividad, Spain (TIN2013-46616-C2-1-R)

Duration: 01-2014 - 12-2017

Main researcher: Dra. Arantza Diaz de Ilarraza and Dr. Koldo Gojenola

Title: QTLeap: QTLeap: Quality Translation by Deep Language Engineering Approaches, FP7-ICT 610516

Institution: European Commission

Duration: 01/11/2013-01/10/2016

Main researcher: Antonio Branco (Universidade de Lisboa)

Title: Antidote

Institution: CHIST-ERA - INT-Acciones de Programación Conjunta Internacional (MINECO) 2020.  
PCI2020-120717-2

Duration: 2021/04/01-2024/03/31

Main researcher: Rodrigo Agerri

### **Research stays**

Center: Center for Cognitive Science (University of Edinburgh)

City: Edinburgh Country: United Kingdom

Year: 1992 Duration: 9 weeks

Theme: Natural Language Processing Contact researcher: Robert Dale

Center: Department of General Linguistics, University of Helsinki

Localidad: Helsinki Country: Finland

Year: 1995 Duration: 2 weeks

Theme: Natural Language Processing, Constraint Grammars, Corpus Processing

Contact researcher: Atro Voutilainen

Center: Department of Informatics, University of Sussex

Localidad: Brighton Country: United Kingdom

Year: 2004 Duration: 10 weeks

Theme: Natural Language Processing, Parsing, Probabilistic approaches to NLP

Contact researcher: John Carroll

Center: Department of Linguistics and Philosophy, University of Uppsala

Localidad: Uppsala Country: Sweden

Year: 2010 Duration: 3 weeks

Theme: Natural Language Processing, Parsing, Dependency Parsing

Contact researcher: Joakim Nivre