http://ixa.eus/

BasqueParl: A Bilingual Corpus of Basque Parliamentary Transcriptions

Nayla Escribano*, Jon Ander González*, Julen Orbegozo-Terradillos, Ainara Larrondo-Ureta**, Simón Peña-Fernández**, Olatz Perez-de-Viñaspre* and Rodrigo Agerri***
* HiTZ Center - Ixa | ** Gureiker (University of the Basque Country UPV/EHU)
{nayla.escribano, olatz.perezdevinaspre, rodrigo.agerri}@ehu.eus

Motivation

- New corpus for **political discourse analysis**
- **Language use difference** in a Basque-Spanish code-switching bilingual corpus
- **Analysis of speaker speech in different categories:** language, gender, age, party…
- **Comparison of parliamentary behaviour** with the society and their expectations

Contributions

- **Release of BasqueParl**:
  - New code-switching bilingual corpus
  - Basque Parliament transcriptions from 2 terms (2012-2016 and 2016-2020)
  - 14 M words
- **Metadata**: gender, birth, party and language
- **Lemmas and named entities**

Paragraph Example

Basque **Spanish**

Gauzak egiten dira eta uste dut nik, nik ere eskubidea dudala Gobernuak eta beste erakundeek egiten dutena esateko. Zeren beti ver el vaso medio vacío o medio lleno, pues cambia un poco la perspectiva y vernos siempre en modo Gobierno, creo que no es nada objetivo. Se hacen cosas, se harán cosas y esta vez creo que me deberían reconocer que de la iniciativa primera a lo que hemos acordado, no nos hemos dejado nada o creo que casi nada. Entonces, bueno, sólo quería aclarar eso eta eskerrak berriro.

→ **Spanish paragraph**

Corpus Description

Field	Field type	Speeches	Paragraphs	Words	Lemmas	Entities
Overall	**Overall**	**41,417**	**342,666**	**13,872,105**	**5,090,573**	**349,890**
Language	Basque (eu)	34,571	133,599	2,938,061	1,375,686	122,439
	Spanish (es)	18,016	209,067	10,934,044	3,714,887	227,451
Gender	Female	30,857	185,119	6,452,503	2,424,765	179,030
	Male	10,559	157,481	7,416,852	2,664,596	170,782
Party	EAJ-PNV	28,530	143,685	4,330,467	1,701,415	145,400
	EH Bildu	3,638	58,391	2,423,245	1,009,715	56,440
	PP	3,562	51,683	2,672,062	858,798	56,376
	PSE-EE	2,834	44,738	2,327,350	769,767	43,994
	EP	1,467	24,235	1,171,948	410,966	24,738
	UPyD	1,368	19,110	909,007	322,220	22,270
Speaker	President	22,772	60,326	753,231	315,138	62,317

Data Analysis

Word: **eu** xor **es** | Paragraph: **eu** xor **es** | Speech: **eu** and/or **es**

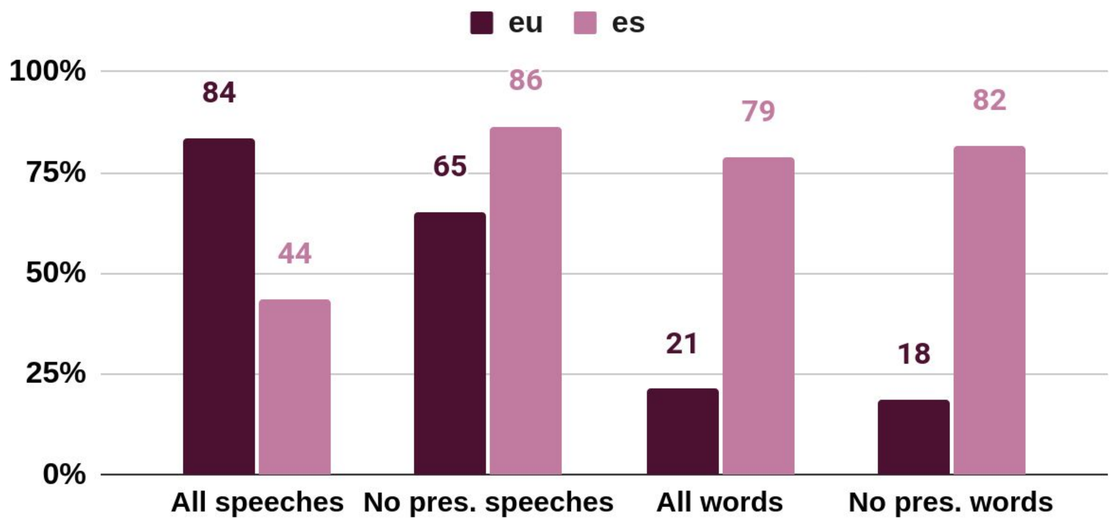

Figure 1: Language use at speech and word level

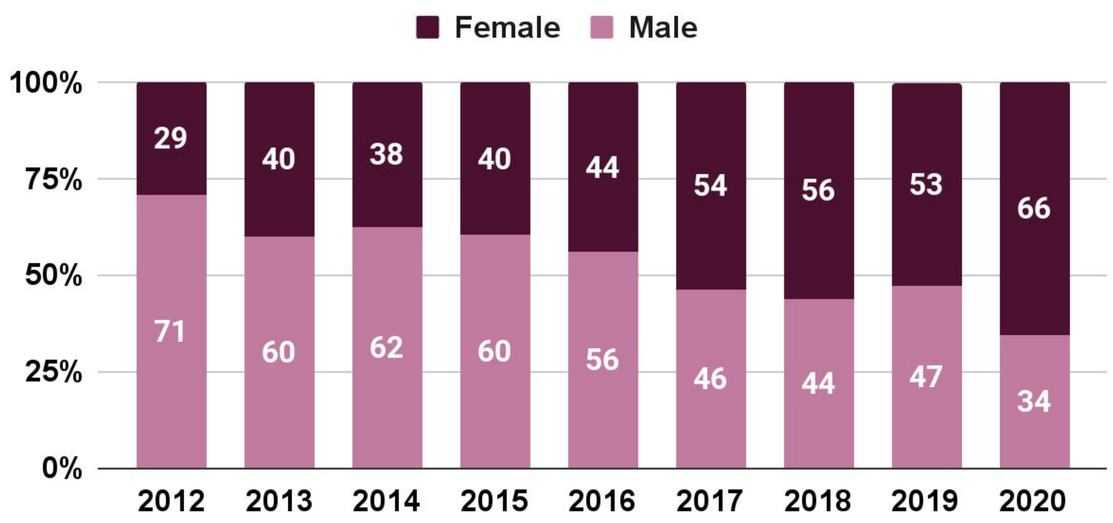

Figure 4: Gender distribution over time at word level

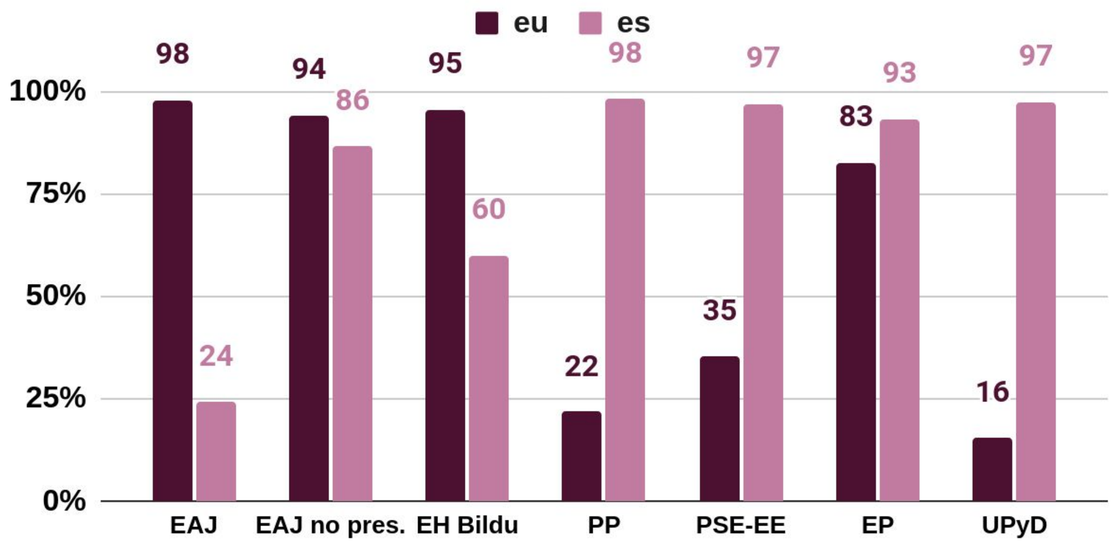

Figure 2: Language use in each party at speech level

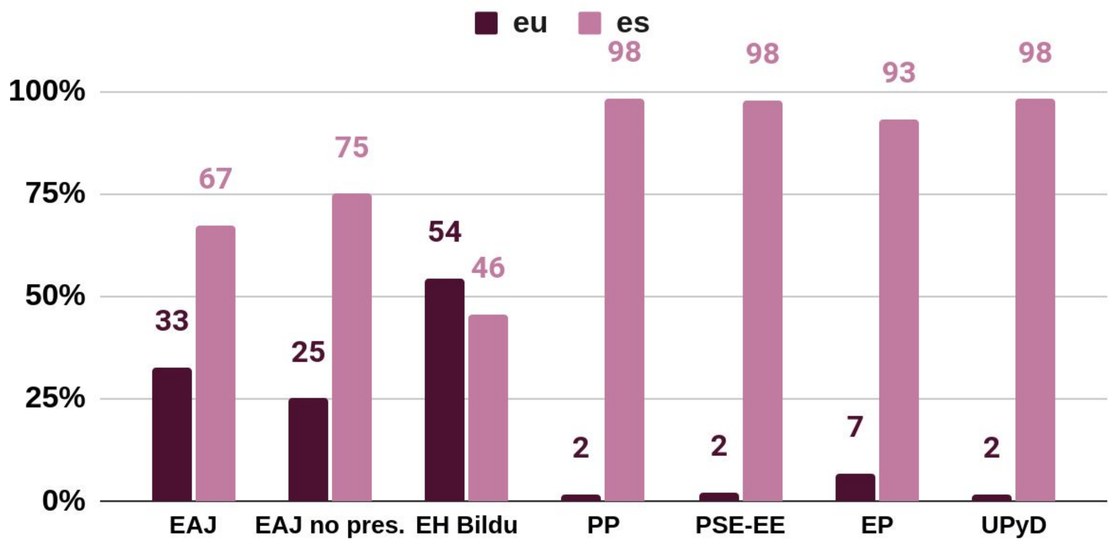

Figure 3: Language use in each party at word level

Concluding Remarks

- **Basque** often used in speeches, but most words in **Spanish**
- Similar results by **party:** only EH Bildu uses more Basque words and some parties' usage is minimal
- **Female** underrepresentation at word level has reversed over the years
- **Release of the corpus** for public research on multilingual and crosslingual NLP tasks and political discourse analysis

Corpus available at: https://github.com/ixa-ehu/basqueparl | Demo: http://legebiltzarra.ixa.eus