Adding the Basque Parliament corpus to ParlaMint project

Workshop on Creating, Enriching and Using Parliamentary Corpora (ParlaCLARIN III at LREC2022)

Jon Alkorta [jon.alkorta@ehu.eus]
Mikel Iruskieta [mikel.iruskieta@ehu.eus]

HiTZ Basque Center for Language Technologies
Ixa, University of the Basque Country (UPV/EHU)

Marseille, 20th June 2022

Outline

Introduction

- The Basque Parliament is one of three parliaments in the Basque Country.
  - Araba, Bizkaia and Gipuzkoa provinces depend on this parliamint.
- Current Basque Parliament:
  - There are 6 political parties: EAJ-PNV, EH Bildu, PSE-EE, Elkarrekin Podemos, PP+CS, Vox.
  - The Basque Government: EAJ-PNV and PSE-EE.
- Languages in the Basque Parliament.
  - Basque and Spanish are official.
  - Not all parliamentarians speak Basque.
  - There is a translation service.
    - At the moment in the parliament.
    - In the acts of parliament.

Outline

Related Works

- Studies related to **voice processing**.
  - [Bordel et al., 2011]: an automatic video subtitling system.
  - **Mintzai-ST** [Etchegoyhen et al., 2021]. This is a Basque-Spanish parallel corpus compiled with both speech and text data.
  - **Euskoparl** [Pérez et al., 2012]. A parallel corpus in Spanish and Basque with both text and speech data.

Related Works

- Studies related to **text processing**.
  - International context: ParlaMint project [Erjavec et al., 2022].
    - Our work is related to this project.
  - **NEW!** BasqueParl [Escribano et al., 2022].
    - Work oriented to text processing.
    - The corpus is not structured.

.

Outline

Aims

General objective

- Add the Basque Parliament Corpus to ParlaMint project.
  - The aims is to create a resource with the Basque Parliament speeches with similar characteristics to other resources of the ParlaMint project.

Secondary objectives

- Create a Basque Parliament Corpus entirely in Basque.
  - The Basque language is under-resourced language, so it is very helpful to create a corpus in Basque with speeches from a specific domain and context.

Outline

1- Collect the parliamentary data

- Make a request to the secretary of the Basque Government.
- The request (2021/1887) was granted on March 21, 2021.

EUSKO LEGEBILTZARRA
PARLAMENTO VASCO

UPV/EHUko irakasle-ikertzailea, Mikel Iruskieta

EUSKO LEGEBILTZARREKO MAHAIAK 2021(E)KO MARTXOAREN 9(E)AN EGINDAKO BILERAN, ERABAKI HAUEK HARTU DITU, BESTEAK BESTE:

Idazkia
Eusko Legebiltzarreko transkripzioak (eta beraien itzulpenak, ahal denean) eskatzen ditu ParlaMint proiektuan euskarazko corpusa egoteko (2021/1887)

- **Egilea**: UPV/EHUko irakasle-ikertzailea, Mikel Iruskieta
- **Norentzat**: Eusko Legebiltzarreko Mahaia

Mahaiak eman du baimena.

Profesor-investigador de la UPV/EHU, Mikel Iruskieta

LA MESA DEL PARLAMENTO VASCO, EN SESIÓN CELEBRADA EL DÍA 9 DE MARZO DE 2021, HA ADOPTADO, ENTRE OTROS, LOS SIGUIENTES ACUERDOS:

Escrito
Solicita las transcripciones del Parlamento Vasco (y sus traducciones, cuando sea posible), para que haya un corpus de euskera en el proyecto ParlaMint (2021/1887)

- **Autor**: Profesor-investigador de la UPV/EHU, Mikel Iruskieta
- **Para**: Mesa del Parlamento Vasco

La Mesa concede la autorización.

2- Create the original version

- The parliamentary data is divided into several files.
- In each DOC file, there are two columns:
  - The left column contains: the original speech.
  - The right column: translation to the other official language.
- We create TEI-XML format document using the left column.

Son los cambios sociales, las nuevas demandas en materia de seguridad, los que hacen necesaria la modificación de la Ley vasca de Policía. Proximidad, conocimiento cercano de la sociedad a la que sirven, compromiso con los valores de la ciudadanía vasca, trabajo coordinado y eficacia policial, son hoy señas de identidad de nuestro modelo policial. Una policía, en definitiva, implicada en una sociedad abierta, plural, bilingüe, constituida por ciudadanas y ciudadanos iguales en derechos.

Gurea bezalako gizarte aurreratu baten segurtasun-premietatik hurbil polizia-zerbitzu eraginkorra eskainiz, herritarron onespen eta konfiantza maila altua lortu du daukagun polizia-ereduak, bai Ertzaintzak eta bai udaltzaingoek.

Gaur aurkezten dugu bosgarren lege-aldaketarako proposamen honekin etorkizunari begiratu nahi diogu, egindakotik eta daukagunetik abiatuta, premia eta erronka berrietara egokitzeko, besteak beste, berrikuntza-prozesu inportante batean gaudelako.

Urte gutxiren buruan milaka ertzain eta udaltzain berri dagoeneko iristen ari dira eta iritsiko dira euskal poliziaren zerbitzuetara, zerbitzari publiko izatera.

Gizarte-aldaketa horiek, segurtasun-alorreko eskakizun berri horiek dira euskal Polizia Legea aldatzeko beharra eragiten dutenak. Hurbiltasuna, zerbitzatu behar duten gizartearen gertuko ezaguera, euskal herritarren balioekiko konpromisoa, lan koordinatua eta polizia-lanaren efikazia dira gaur egun gure polizia-ereduaren nortasun-ezaugarriak. Laburbilduz, polizia txertatuta dago irekia, askotarikoa eta elebiduna den eta eskubide-berdintasuna duten herritarrez osatuta dagoen gizartearekin.

Al ofrecer un servicio policial eficaz y cercano a las necesidades de seguridad de una sociedad avanzada como la nuestra, nuestro modelo policial –tanto la Ertzaintza como las policías locales– ha alcanzado un alto grado de reconocimiento y confianza por parte de la ciudadanía.

Con esta propuesta de quinta modificación legal, queremos mirar hacia el futuro, partiendo de lo ya realizado y existente, para adaptarnos a las nuevas necesidades y retos, entre otras cuestiones, porque nos hallamos inmersos en un importante proceso de renovación.

En el plazo de pocos años, miles de ertzainas y policías locales están accediendo y van a seguir incorporándose a los servicios policiales vascos, para convertirse en servidores públicos.

Irudia 2: The file with two columns.

```
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg175" lang="es">Creo que estamos en un momento político un tanto… no sé si convulso o cómo decirlo, pero creo que
no debiéramos utilizar la tribuna únicamente pensando en las siglas de los partidos políticos en los que estamos, porque creo que Osakidetza está por encima de lo que
tengamos que decir todos. Por tanto, somos partidos políticos todos responsables, y no utilicemos determinadas situaciones pensando mucho más en réditos electorales,
réditos partidistas. Aquellos que dicen que solo se preocupan por la institución, por los médicos, por el dinero de los contribuyentes, por la mejor gestión…, lejos de
eso, lo único que están haciendo ustedes es perjudicar a la institución que estaba mejor valorada por los ciudadanos vascos.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u22" who="#TejeriaOtermin" ana="#chair">
          <note>LEHENDAKARIAK:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg176" lang="eu">Eskerrik asko, Rojo anderea.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg177" lang="eu">Euzko Abertzaleak taldearen ordezkaria. Larrauri anderea, zurea da hitza.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u23" who="#LarrauriAranguren">
          <note>LARRAURI ARANGUREN andreak:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg178" lang="eu">Bai, eskerrik asko, legebiltzarburu anderea. Sailburuok</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg179" lang="eu">-bereziki, Murga anderea-, legebiltzarkideok, jaun-andreok, egun on.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg180" lang="es">Bueno, hoy debatimos una moción consecuencia de interpelación que se basa en una denuncia de un
sindicato -el sindicato LAB-, que parece que el grupo proponente hace propia, y, en concreto, se refiere a la gestión del contrato de hostelería de la OSI Araba, contrato
que incluye los servicios de cocina de ambos hospitales, Santiago y Txagorritxu, además del servicio al hospital psiquiátrico de Araba y el servicio de cafetería. Y, en
concreto, la denuncia se centra en la cocina del Hospital Santiago, que no sé si…, o sea, la señora Ubera ha dicho que la más pringada. La verdad es que me parece una… No
sé si he entendido bien, pero me parece una calificación bastante poco apropiada e irrespetuosa, pero, bueno. Y está claro…, totalmente de acuerdo en que Osakidetza
gestiona dinero público, pero, bueno, a estas alturas, hacer esas afirmaciones… Yo creo que ninguno tenemos ninguna duda de que, desde luego, pues, debe primar la buena
gestión.</seg>
```


Irudia 3: The original corpus.

3- Create the Basque version

- We choose the passages written in Basque from:
  - The left column (original text).
  - The right column (translated text).
- The script first calculates how likely the paragraphs are to be in Basque or Spanish.
- The script takes the paragraphs most likely to be written in Basque from the left column or from the right column.

```
          <note>LEHENDAKARIAK:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg271" lang="eu">Beraz, ez da onartu.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg272" lang="eu">Ez denez onartu, jarraian bozkatuko dugu Euskal Talde Popularraren osoko zuzenketa. Bozkatu
dezakegu.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg273" lang="eu">Botazioa eginda, hauxe izan da emaitza: emandako botoak, 74; aldekoak, 9; aurkakoak, 48;
abstentzioak, 17.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u37" who="#TejeriaOtermin" ana="#chair">
          <note>LEHENDAKARIAK:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg274" lang="eu">Beraz, ez da onartu.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg275" lang="eu">Eta, azkenik, Euskal Sozialistak eta Euzko Abertzaleak taldeek aurkeztu duten osoko zuzenketa
bozkatuko dugu. Bozkatu dezakezue.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg276" lang="eu">Botazioa eginda, hauxe izan da emaitza: emandako botoak, 74; aldekoak, 37; aurkakoak, 37;
abstentzioak, 0.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u38" who="#TejeriaOtermin" ana="#chair">
          <note>LEHENDAKARIAK:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg277" lang="eu">Berdinketa dagoenez, berriro bozkatuko dugu. Bozkatu dezakezue.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg278" lang="eu">Botazioa eginda, hauxe izan da emaitza: emandako botoak, 74; aldekoak, 37; aurkakoak, 37;
abstentzioak, 0.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u39" who="#TejeriaOtermin" ana="#chair">
          <note>LEHENDAKARIAK:</note>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg279" lang="eu">Beraz, berdinketa dagoenez, hurrengo osoko bilkurarako utziko dugu.</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg280" lang="eu">Gai-zerrendako bosgarren puntua: "Mozioa, Carmelo Barrio Baroja Euskal Talde Popularreko
legebiltzarkideak aurkeztua, Eusko Jaurlaritzak Kultura Ondarearen Euskal Autonomia Erkidegoko Kontseiluaren osaerari, antolamenduari, funtzionamenduari eta zereginei
dagokienez dituen irizpideen inguruan. Eztabaida eta behin betiko ebazpena".</seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg281" lang="eu">Talde proposatzailearen txanda. Barrio jauna, zurea da hitza ekimena aurkeztu eta defendatzeko.</
seg>
          <seg id="ParlaMint-EU-2020-02-06zenb1392-1399.seg282" lang="eu">Mesedez, isiltasuna.</seg>
        </u>
        <u id="ParlaMint-EU-2020-02-06zenb1392-1399.u40" who="#BarrioBaroja">
```

Irudia 4: The Basque corpus.

4- Metadata file

- The metadata file is valid for both versions of the corpus.
- The root file is in Basque, Spanish and English.
  - The title, the size, the date of creation of the corpus, as well as political parties, parliamentarians, sessions and positions

```
<person xml:id="PinedoBustamante">                                          <!-- ##58 -->
   <persName>
      <surname>Pinedo</surname>
     <surname>Bustamante</surname>
      <forename>Leire</forename>
   </persName>
  <sex value="F">emakume</sex>
   <birth when="1976">
      <placeName ref="Sopela"></placeName>
   </birth>
   <affiliation role="member" ref="EHBildu" from="2007" ana="#DZ.8"/>
   <idno type="wikimedia" xml:lang="eu">https://eu.wikipedia.org/wiki/Leire_Pinedo</idno>
</person>
<person xml:id="PrietoSanVicente">                                          <!-- ##59 -->
   <persName>
      <surname>Prieto</surname>
     <surname>San Vicente</surname>
      <forename>Txarli</forename>
   </persName>
  <sex value="M">gizon</sex>
   <birth when="1957-03-29">
      <placeName ref="Gasteiz"></placeName>
   </birth>
   <affiliation role="member" ref="PSE-EE" from="1991-06-17" ana="#DZ.8"/>
   <idno type="wikimedia" xml:lang="eu">https://eu.wikipedia.org/wiki/Txarli_Prieto</idno>
</person>
```

Irudia 5: The metadate file.

Outline

Characteristics

- The texts obtained are from the speeches between February, 2015 and February, 2021.

	Basque corpus	**Original corpus**
Basque	7.37	1.98
Spanish		7.35
Unidentified		0.05
Total	**7.37**	**9.38**

Taula 1: Estimation of size of the corpus in words

(in millions)

Characteristics II. Unidentified words

- Several for various reasons, the language could not be identified:
  - Words from other languages.
  - The script has given the same probability to the word to be Basque or Spanish.

Hypothesis and discussion

- Sociolinguistic Study. It is used Spanish more...
  - when parliamentarians want to say something important?
  - when parliamentarians want to express emotions?
- Other Possible Studies.
  - Text complexity: original texts vs. translated texts.

Outline

- We present two versions of the Basque Parliament Corpus.
- Future Works:
  - Adapt the corpus to the ParlaMint format.
  - The linguistic processing following the guidelines of the ParlaMint project.
  - Study some hypotheses mentioned before.

Adding the Basque Parliament corpus to ParlaMint project

Workshop on Creating, Enriching and Using Parliamentary Corpora (ParlaCLARIN III at LREC2022)

Jon Alkorta [jon.alkorta@ehu.eus]
Mikel Iruskieta [mikel.iruskieta@ehu.eus]

HiTZ Basque Center for Language Technologies
Ixa, University of the Basque Country (UPV/EHU)

Marseille, 20th June 2022

References I


Bordel, G., Nieto, S., Penagarikano, M., Rodriguez-Fuentes, L. J., and Varona, A. (2011).
Automatic subtitling of the basque parliament plenary sessions videos.
In *Twelfth Annual Conference of the International Speech Communication Association*.


Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al. (2022).
The parlamint corpora of parliamentary proceedings.
Language resources and evaluation, pages 1–34.

References II


Escribano, N., González, J. A., Orbegozo-Terradillos, J., Larrondo-Ureta, A., Peña-Fernández, S., Perez-de Viñaspre, O., and Agerri, R. (2022).
Basqueparl: A bilingual corpus of basque parliamentary transcriptions.
arXiv preprint arXiv:2205.01506.


Etchegoyhen, T., Arzelus, H., Ugarte, H. G., Alvarez, A., González-Docasal, A., and Fernandez, E. B. (2021).
Mintzai-ST: Corpus and baselines for Basque-Spanish speech translation.
Proc. IberSPEECH 2021, pages 190–194.

References III


Pérez, A., Alcaide, J. M., and Torres, M.-I. (2012).
EuskoParl: a speech and text Spanish-Basque parallel corpus.
In *Thirteenth Annual Conference of the International Speech Communication Association*.