

ParlaMint 2.0: un proyecto de CLARIN-K

Castellano, gallego, euskera y catalán

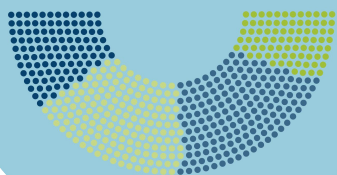
Jon Alkorta* & Elisa Fernández Rei**

*Universidad del País Vasco/Euskal Herriko Unibertsitatea,

**Instituto da Lingua Galega, Universidade de Santiago de Compostela

jon.alkorta@ehu.eus, elisa.fernandez@usc.gal

ParlaMint

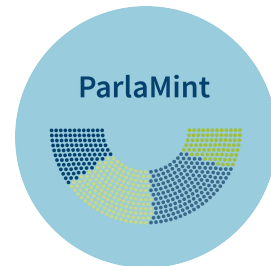


Madrid, 13 de septiembre de 2022



Introducción

CLARIN



- ParlaMint es un proyecto, apoyado financieramente por CLARIN ERIC, que contribuye a la creación de corpus multilingües de textos parlamentarios, comparables y anotados uniformemente.
- Proporciona datos que permiten el análisis paneuropeo del discurso parlamentario desde distintas perspectivas y disciplinas.
- Fases:
 - ParlaMint I (julio 2020-mayo 2021): publicación de corpus de 17 lenguas (alrededor de 500 millones de palabras)
 - ParlaMint II (diciembre 2021-mayo 2023): actualización de los datos existentes, mejora de la usabilidad de los corpus disponibles e incorporación de 10 nuevas lenguas

Objetivos

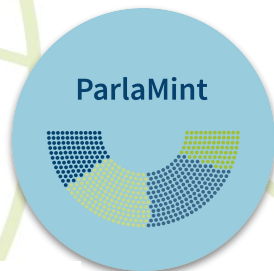
1. Ofrecer datos de los distintos parlamentos a través de recursos que permitan su comparación
2. Proporcionar estos datos y recursos de modo que sean comprensibles para su uso por parte de investigadores, periodistas, ciudadanía en general...

Query covid 123 (5.21 per million)

Document ID	Snippet
2020-07-13,ČrnaZvonko	obstoječih sredstev za potrebe covid . To smo tudi naredili in omogočili
2020-07-13,CiglerVrajšJanez	pristnosti po koncu prvega vala epidemije covid . Kot rečeno, veliko naših obveznosti je
2020-07-13,ŠircaAndrej	oziroma so za toliko večji, in to zaradi krize covid . Te ocene, o katerih ste govorili, so ne prvi
2020-07-13,ŠircaAndrej	, kako bo (LOC: Slovenija) / ob tem, da seveda ne bo več covid krize, pršla v stanje, ko bodo prihodili in
2020-07-13,GantarTomaž	za starejši problem tudi sicer, ne samo okužbe covid , vsako leto tudi mnogi zbolijo in tudi lahko
2020-07-13,KustecLipicerSimona	nacionalnih športnih federacij, in zdaj na covid vprašanje - Kaj smo naredili s tremi
2020-07-13,KustecLipicerSimona	zapolnili preko ostalih oblik, ki jih je covid nekoliko zaobjel, kot pomoč v času, ko niso mogli
2020-07-13,KovačičKaj	upoštevati nelihih novih oblik, ki so prilele s tem covidom, me zanima tudi: kaj kaka je vaša ocena
2020-07-13,KustecLipicerSimona	covid situacije in na delovanja družtev. V času covid niti ne deloval, tudi običajnih stroškov
2020-07-13,Prebilnik	evropsko kohezijsko politiko in njen razvoj po covidu oziroma med covidom ; nekako o temi, ki je za nekoga državljana, ki
2020-07-13,Prebilnik	politiko in njen razvoj po covidu oziroma med covidom ; nekako o temi, ki je za nekoga državljana, ki
2020-07-13,ČrnaZvonko	tematskih ciljev je bil pogojen z ukrepi covid in v tem delu je tudi odstopanje mogoče.
2020-07-13,ČrnaZvonko	, da se na posameznih oseh ne glede na situacijo covid pojavljajo problemi, so se pojavljali že do
2020-07-13,ČrnaZvonko	odstopanje zaradi preizkušenosti za potrebe covid , tako kot določajo veljavne uredbe iz marca in
2020-07-10,StarovičVojko	nabave mask in respiratorjev med epidemijo COVID . Kako ironično, da je ustanoviteljica
2020-05-26,TomičVioleta	bi poslanci od njih pričakovali. V redu, zaradi covid smo se izogibali [gop:inaudible: [-]] Počakajte, da končam, ne
2020-05-30,TomičVioleta	tzjemno nesposobnost. Žal, če ne bi bilo krize s covidom , bi si jo ta vlada morala izmysliti, kajti njene
2020-05-14,TomičVioleta	za boljšo voljo želeli smo sodelovati že pri /MISC COVID 1 / in COVID 2 zakonu pa smo videli kako ste nas
2020-05-14,TomičVioleta	voljo želeli smo sodelovati že pri /MISC COVID 1 / in COVID 2 zakonu pa smo videli kako ste nas zaobčili in
2020-05-14,TomičMateja	opreme zagotovili, potem nebi mogli zdraviti COVID bolnikov, ker so dejali, da nimajo srca, da bi v

Page 1 of 7 Go Next Last

Ej. [Parlamento esloveno](#)



Preguntas de investigación

1. ¿Qué temas son más frecuentes en cada momento en los distintos parlamentos?
¿Cómo cambian?
2. ¿Qué variables sociolingüísticas condicionan la forma de hablar de los parlamentarios?
3. ¿Cómo se construyen discursivamente las posiciones de izquierda/centro/derecha?
4. ¿Qué mecanismos lingüísticos y discursivos se usan para construir las posiciones de gobierno vs. oposición?
5. Los grupos afines ideológicamente (populares, verdes, socialistas...) ¿Hablan de la misma manera y sobre los mismos temas en los diferentes parlamentos?
6. En los contextos bilingües y multilingües, ¿qué factores condicionan la elección de lengua o lenguas por parte de los políticos?

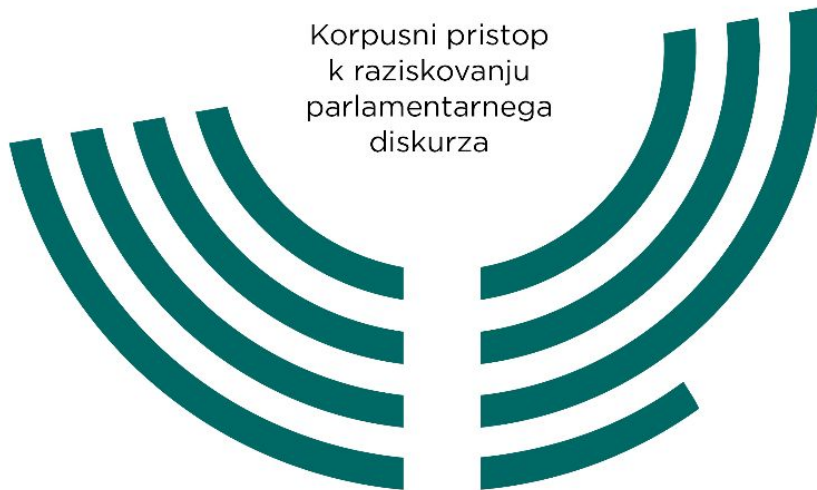
Materiales para el uso práctico del corpus

- 1 Introduction
- 2 Instructions for use
- 3 Corpora and concordancers
 - 3.1 Corpora
 - 3.2 Concordancers
- 4 Parliamentary records
 - 4.1 Parliamentary discourse
 - 4.2 Faithfulness of the records
 - 4.3 Know your research dataset
- 5 Language and gender
- 6 Corpus analysis
 - 6.1 The siParl 2.0 corpus
 - 6.2 TASK 1: Representation of women in the Slovenian Parliament
 - 6.2.1 Creating subcorpora
 - 6.2.2 Using frequency lists
 - 6.2.3 Comparative analysis
 - 6.3 TASK 2: Issues addressed by women
 - 6.3.1 Extracting keywords
 - 6.3.2 Analysing concordances
 - 6.3.3 Comparative analysis
 - 6.4 TASK 3: Topics related to women
 - 6.4.1 Working with frequencies
 - 6.4.2 Extracting collocations
 - 6.4.3 Comparative analysis

Voices of the Parliament A Corpus Approach to Parliamentary Discourse Research

»Prvič, sem političarka in
ne politik, drugič pa ...«

Korpusni pristop
k raziskovanju
parlamentarnega
diskurza



Recursos para la explotación del corpus

- Los corpus ParlaMint se han incorporado a SketchEngine:
<https://www.sketchengine.eu/parlamint-corpora-of-parliamentary-debates/>

Tools to work with the ParlaMint corpora

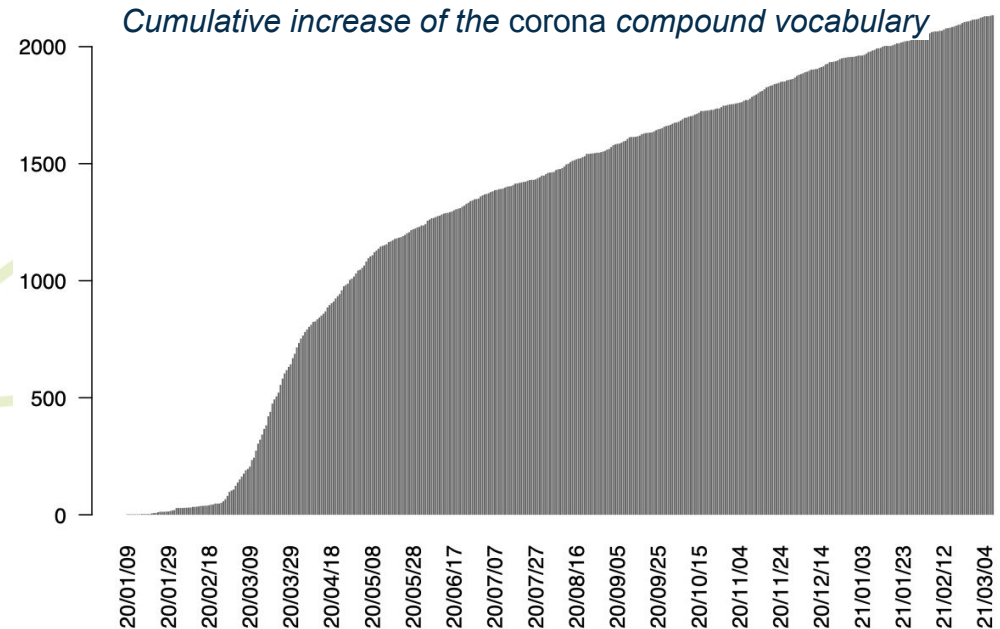
A complete set of Sketch Engine tools is available to work with these parliamentary corpora to generate:

- **word sketch** – collocations categorized by grammatical relations
- **thesaurus** – synonyms and similar words for every word
- **keywords** – terminology extraction of one-word and multi-word units
- **word lists** – lists of nouns, verbs, adjectives etc. organized by frequency
- **n-grams** – frequency list of multi-word units
- **concordance** – examples in context
- **text type analysis** – statistics of metadata in the corpus




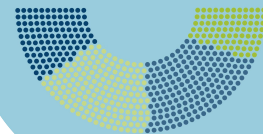
Casos de uso de impacto

- Corpus: Norwegian Newspaper Corpus (en CLARIN Resource Family 'Newspaper Corpora')
- Herramienta para la gestión de corpus y motor de búsqueda: **Corpuscle**, desarrollado en CLARINO Bergen Centre



Procesado de datos

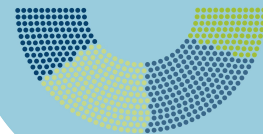
1. Obtener datos del parlamento y sus metadatos
 2. **Convertirlos al esquema de ParlaMint**
 3. **Anotación lingüística (UDpipe y NERC)**
- 
- The slide features several decorative, overlapping curved lines in a light green color that sweep across the bottom and right sides of the page. A solid green horizontal bar is located at the very bottom of the slide.



- Ejemplo de un fichero de metadatos:

```
<taxonomy xml:id="speaker_types">
  <desc xml:lang="eu"><term>Hiztun motak</term></desc>
  <desc xml:lang="en"><term>Types of speakers</term></desc>
  <category xml:id="chair">
    <catDesc xml:lang="eu"><term>Mahaiburua</term>: bilera bateko burua</catDesc>
    <catDesc xml:lang="en"><term>Chairperson</term>: chairman of a meeting</catDesc>
  </category>
  <category xml:id="regular">
    <catDesc xml:lang="eu"><term>Erregularra</term>: bilera bateko ohiko hizlaria</catDesc>
    <catDesc xml:lang="en"><term>Regular</term>: a regular speaker at a meeting</catDesc>
  </category>
  <category xml:id="guest">
    <catDesc xml:lang="eu"><term>Gonbidatua</term>: bilera bateko hizlari gonbidatua</catDesc>
    <catDesc xml:lang="en"><term>Guest</term>: a guest speaker at a meeting</catDesc>
  </category>
</taxonomy>
```

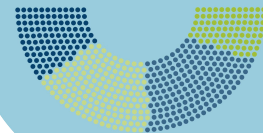
Taxonomía de las intervenciones



- Ejemplo de un fichero de metadatos:

```
<!-- #5 -->
<person xml:id="AlonsoMartín">
  <persName>
    <surname>Alonso</surname>
    <surname>Martín</surname>
    <forename>Alberto</forename>
  </persName>
  <sex value="M">gizon</sex>
  <birth when="1976">
    <placeName ref="">Oñati</placeName>
  </birth>
  <affiliation role="member" ref="PSE-EE" from="2020-07-12" ana="#DZ.8"/>
  <idno type="wikimedia" xml:lang="eu">https://eu.wikipedia.org/wiki/Alberto_Alonso_Mart%C3%ADn</idno>
</person>
<!-- #6 -->
<person xml:id="ÁlvarezMartínez">
  <persName>
    <surname>Álvarez</surname>
    <surname>Martínez</surname>
    <forename>Gorka</forename>
  </persName>
  <sex value="M">gizon</sex>
  <birth when="1988-08-02">
    <placeName ref="">Sestao</placeName>
  </birth>
  <affiliation role="member" ref="EAJ-PNV" from="2015-06-13" ana="#DZ.8"/>
  <idno type="wikimedia" xml:lang="spa">https://es.wikipedia.org/wiki/Gorka_%C3%81lvarez</idno>
</person>
```

Interventores



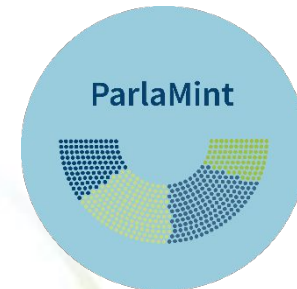
- Ejemplo de un fichero de metadatos:

```
<org xml:id="EAJ/PNV" role="political_party">
  <orgName full="yes" xml:lang="eu">Euzko Alderdi Jeltzalea</orgName>
  <orgName full="yes" xml:lang="en">Basque Nationalist Party</orgName>
  <orgName full="init">EAJ/PNV</orgName>
  <event from="1895-07-31">
    <label xml:lang="en">existence</label>
  </event>
  <idno type="wikimedia" xml:lang="eu">https://eu.wikipedia.org/wiki/Euzko_Alderdi_Jeltzalea</idno>
  <idno type="wikimedia" xml:lang="en">https://en.wikipedia.org/wiki/Basque_Nationalist_Party</idno>
</org>
<!-- ##1- EAJ/PNV -->

<org xml:id="EHBildu" role="political_party">
  <orgName full="yes" xml:lang="eu">Euskal Herria Bildu</orgName>
  <orgName full="yes" xml:lang="en">Euskal Herria Bildu</orgName>
  <orgName full="init">EH Bildu</orgName>
  <event from="2012-06-12">
    <label xml:lang="en">existence</label>
  </event>
  <idno type="wikimedia" xml:lang="eu">https://eu.wikipedia.org/wiki/Euskal_Herria_Bildu</idno>
  <idno type="wikimedia" xml:lang="en">https://en.wikipedia.org/wiki/EH_Bildu</idno>
</org>
```

Partidos políticos

ParlaMint-GALICIA: exemplo



A señora PRESIDENTA: Grupo Mixto, señora Iglesias e señora Martínez.

A señora IGLESIAS SUEIRO: Bos días, señora conselleira; bos días a todas. Grazas, presidenta.

De todas é coñecido o caos rampante nas Urxencias nas últimas datas, que coinciden co andazo da gripe: esperas de ata sete horas, profesionais desbordadas, escaseza de camas e de material...

E pese á petición de perdón por parte do presidente, coma se dun pecado se tratara e non dunha responsabilidade política, e dos argumentos desta conselleira, que reduce a situación caótica a un problema de confort, señora conselleira, -señora conselleira, ¡que falta de respecto!-...

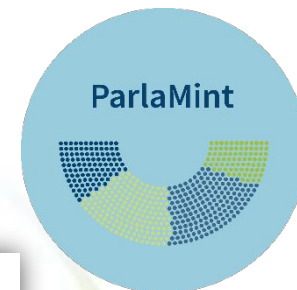
A señora PRESIDENTA: Señora Iglesias.

A señora IGLESIAS SUEIRO: ...e das poucas disimuladas intencións de culpabilizar as pacientes (*Murmurios.*), pois parece que enfermamos por riba das nosas posibilidades...

Conversión TXT-TEI

```
3002 </u>
3003 <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u61">
3004 <seg xml:id="DSPG_95_27012015.seg448"> Grupo Mixto, señora Iglesias e señora Martínez.</seg>
3005 </u>
3006 <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u62">
3007 <seg xml:id="DSPG_95_27012015.seg449"> Bos días, señora conselleira; bos días a
3008 todas. Grazas, presidenta.</seg>
3009 <seg xml:id="DSPG_95_27012015.seg450">De todas é coñecido o caos rampante nas Urxencias nas últimas datas,
3010 que coinciden co andazo da gripe: esperas de ata sete horas,
3011 profesionais desbordadas, escaseza de camas e de material... </seg>
3012 <seg xml:id="DSPG_95_27012015.seg451">E pese á petición de perdón por parte do presidente, coma se dun
3013 pecado se tratara e non dunha responsabilidade política, e dos
3014 argumentos desta conselleira, que reduce a situación caótica a un
3015 problema de confort, señora conselleira, -señora conselleira, ¡que
3016 falta de respecto!-...</seg>
3017 </u>
3018 <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u63">
3019 <seg xml:id="DSPG_95_27012015.seg452"> Señora Iglesias.</seg>
3020 </u>
3021 <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u64">
3022 <seg xml:id="DSPG_95_27012015.seg453">
3023 ...e das poucas disimuladas intencións de
3024 culpabilizar as pacientes
3025 <note>(Murmurios.)</note>
```

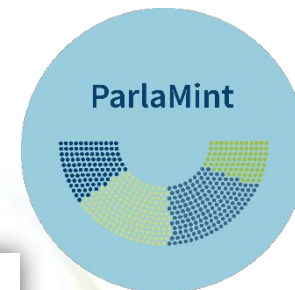
ParlaMint-GALICIA: exemplo



Identificación de interventor

```
3002      </u>
3003      <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u61">
3004          <seg xml:id="DSPG_95_27012015.seg448"> Grupo Mixto, señora Iglesias e señora Martínez.</seg>
3005      </u>
3006      <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u62">
3007          <seg xml:id="DSPG_95_27012015.seg449"> Bos días, señora conselleira; bos días a
3008 todas. Grazas, presidenta.</seg>
3009          <seg xml:id="DSPG_95_27012015.seg450">De todas é coñecido o caos rampante nas Urxencias nas últimas datas,
3010 que coinciden co andazo da gripe: esperas de ata sete horas,
3011 profesionais desbordadas, escaseza de camas e de material... </seg>
3012          <seg xml:id="DSPG_95_27012015.seg451">E pese á petición de perdón por parte do presidente, coma se dun
3013 pecado se tratara e non dunha responsabilidade política, e dos
3014 argumentos desta conselleira, que reduce a situación caótica a un
3015 problema de confort, señora conselleira, -señora conselleira, ¡que
3016 falta de respecto!-...</seg>
3017      </u>
3018      <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u63">
3019          <seg xml:id="DSPG_95_27012015.seg452"> Señora Iglesias.</seg>
3020      </u>
3021      <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u64">
3022          <seg xml:id="DSPG_95_27012015.seg453">
3023              ...e das poucas disimuladas intencións de
3024 culpabilizar as pacientes
3025          <note>(Murmurios.)</note>
```

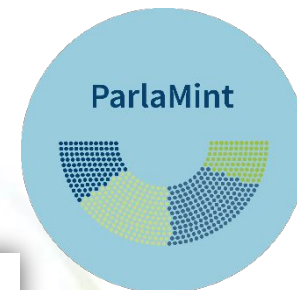
ParlaMint-GALICIA: exemplo



Identificación de rol

```
3002         </u>
3003         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u61">
3004             <seg xml:id="DSPG_95_27012015.seg448"> Grupo Mixto, señora Iglesias e señora Martínez.</seg>
3005         </u>
3006         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u62">
3007             <seg xml:id="DSPG_95_27012015.seg449"> Bos días, señora conselleira; bos días a
3008 todas. Grazas, presidenta.</seg>
3009             <seg xml:id="DSPG_95_27012015.seg450">De todas é coñecido o caos rampante nas Urxencias nas últimas datas,
3010 que coinciden co andazo da gripe: esperas de ata sete horas,
3011 profesionais desbordadas, escaseza de camas e de material... </seg>
3012             <seg xml:id="DSPG_95_27012015.seg451">E pese á petición de perdón por parte do presidente, coma se dun
3013 pecado se tratara e non dunha responsabilidade política, e dos
3014 argumentos desta conselleira, que reduce a situación caótica a un
3015 problema de confort, señora conselleira, -señora conselleira, ¡que
3016 falta de respecto!-...</seg>
3017         </u>
3018         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u63">
3019             <seg xml:id="DSPG_95_27012015.seg452"> Señora Iglesias.</seg>
3020         </u>
3021         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u64">
3022             <seg xml:id="DSPG_95_27012015.seg453">
3023                 ...e das poucas disimuladas intencións de
3024 culpabilizar as pacientes
3025             <note>(Murmurios.)</note>
```

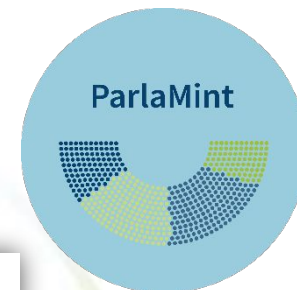
ParlaMint-GALICIA: exemplo



Inicio y fin de intervención y segmentación por párrafos

```
3002         </u>
3003         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u61">
3004             <seg xml:id="DSPG_95_27012015.seg448"> Grupo Mixto, señora Iglesias e señora Martínez.</seg>
3005         </u>
3006         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u62">
3007             <seg xml:id="DSPG_95_27012015.seg449"> Bos días, señora conselleira; bos días a
3008 todas. Grazas, presidenta.</seg>
3009             <seg xml:id="DSPG_95_27012015.seg450">De todas é coñecido o caos rampante nas Urxencias nas últimas datas,
3010 que coinciden co andazo da gripe: esperas de ata sete horas,
3011 profesionais desbordadas, escaseza de camas e de material... </seg>
3012             <seg xml:id="DSPG_95_27012015.seg451">E pese á petición de perdón por parte do presidente, coma se dun
3013 pecado se tratara e non dunha responsabilidade política, e dos
3014 argumentos desta conselleira, que reduce a situación caótica a un
3015 problema de confort, señora conselleira, -señora conselleira, ¡que
3016 falta de respecto!-...</seg>
3017         </u>
3018         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u63">
3019             <seg xml:id="DSPG_95_27012015.seg452"> Señora Iglesias.</seg>
3020         </u>
3021         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u64">
3022             <seg xml:id="DSPG_95_27012015.seg453">
3023                 ...e das poucas disimuladas intencións de
3024 culpabilizar as pacientes
3025             <note>(Murmurios.)</note>
```

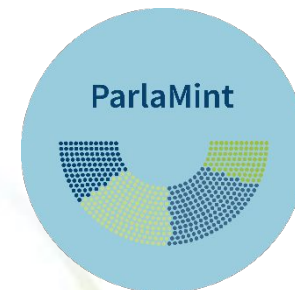
ParlaMint-GALICIA: exemplo



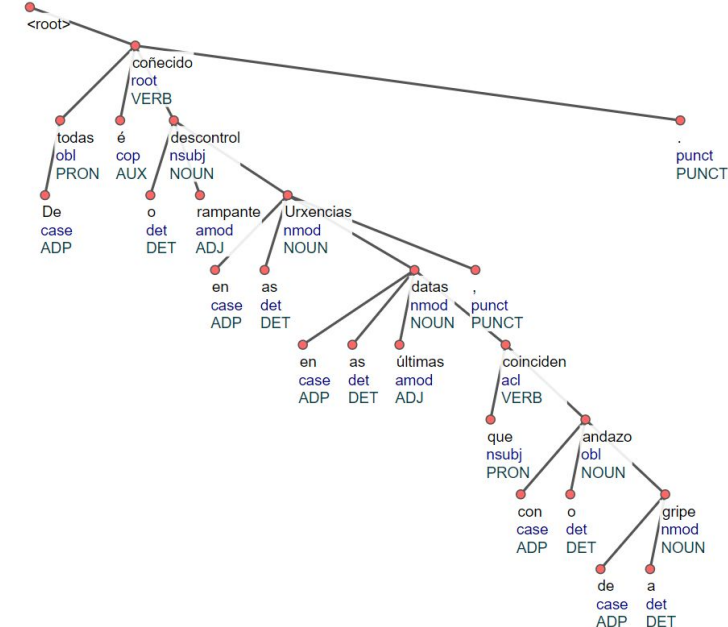
Nota de transcripción

```
3002         </u>
3003         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u61">
3004             <seg xml:id="DSPG_95_27012015.seg448"> Grupo Mixto, señora Iglesias e señora Martínez.</seg>
3005         </u>
3006         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u62">
3007             <seg xml:id="DSPG_95_27012015.seg449"> Bos días, señora conselleira; bos días a
3008 todas. Grazas, presidenta.</seg>
3009             <seg xml:id="DSPG_95_27012015.seg450">De todas é coñecido o caos rampante nas Urxencias nas últimas datas,
3010 que coinciden co andazo da gripe: esperas de ata sete horas,
3011 profesionais desbordadas, escaseza de camas e de material... </seg>
3012             <seg xml:id="DSPG_95_27012015.seg451">E pese á petición de perdón por parte do presidente, coma se dun
3013 pecado se tratara e non dunha responsabilidade política, e dos
3014 argumentos desta conselleira, que reduce a situación caótica a un
3015 problema de confort, señora conselleira, -señora conselleira, ¡que
3016 falta de respecto!-...</seg>
3017         </u>
3018         <u who="RojoPilarMilagros" ana="chair" xml:id="DSPG_95_27012015.u63">
3019             <seg xml:id="DSPG_95_27012015.seg452"> Señora Iglesias.</seg>
3020         </u>
3021         <u who="IglesiasCarmen" ana="regular" xml:id="DSPG_95_27012015.u64">
3022             <seg xml:id="DSPG_95_27012015.seg453">
3023                 ...e das poucas disimuladas intencións de
3024 culpabilizar as pacientes
3025                 <note>(Murmurios.)</note>
```


Anotación lingüística



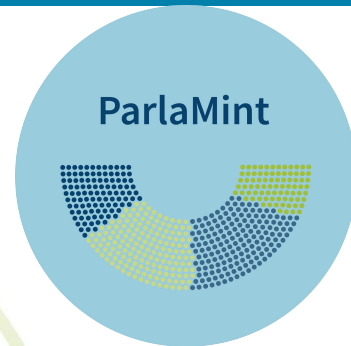
De todas é coñecido o descontrol rampante en as Urxencias en as últimas datas , que coinciden con o andazo de a gripe .



- Segmentación y tokenización
- Lematización y categorías gramaticales
- **Anotación gramatical (*Universal Dependencies*)**
- Reconocimiento de entidades nombradas (*NER*)

<http://ufal.mff.cuni.cz/udpipe>

ParlaMint-ES



- **Participantes:**

- María Calzada es coordinadora de <http://blogs.uji.es/ecpc/> y del corpus en castellano de ParlaMint.
- Presentación en el Webinar de INTELE: <https://youtu.be/b0oNElZbV9E>

- **Características del corpus:**

- Colección de sesiones parlamentarias entre 01-2015 y 12-2020.
- En total, 13,10 millones de palabras.
 - 2,19 millones de palabras por año.
- Más información:
 - [<https://link.springer.com/article/10.1007/s10579-021-09574-0>].



ParlaMint-ES

ParlaMint

- A diferencia de otros idiomas, el castellano está incluido en la primera versión de ParlaMint.
 - El corpus ya está disponible a través de programas de concordancia (noSketch Engine [<https://www.clarin.si/noske/parlamint.cgi/>] / KonText) y ParlaMeter

The screenshot displays the ParlaMint-ES search interface. At the top, the search bar contains the query 'covid' and the engine name 'ParlaMint-SI 2.0 (Slovenian parliament)'. Below the search bar, the results are displayed in a table format. The first column shows the date of the document, the second column shows the document title, and the third column shows the search results for the query 'covid'. The search results are highlighted in yellow. The interface also includes a sidebar with navigation options such as Home, Search, Word list, Corpus info, My jobs, User guide, Save, Make subcorpus, View options, KWIC, Sentence, Sort, Left, Right, Node, References, Shuffle, Sample, Filter, Sub-hits, 1st hit in doc, Frequency, Node forms, Doc IDs, Text types, Collocations, and Visualize.

Date	Document Title	Search Results
2020-07-13	ČrnaZvonko	obstoječih sredstev za potrebe covid . To smo tudi naredili in omogočili
2020-07-13	CiglerKrajJ.Janez	pristojnosti po koncu prvega vala epidemije covid . Kot rečeno, veliko naših obveznosti je
2020-07-13	ŠirceJAndrej	oziroma so za toliko večji, in to zaradi krize covid . Te ocene, o katerih ste govorili, so na prvi
2020-07-13	ŠirceJAndrej	, kako bo /LOC: Slovenija / ob tem, da seveda ne bo več covid . krize, prišla v stanje, ko bodo prihodili in
2020-07-13	GantarTomaž	za starejši problem tudi sicer, ne samo okužbe covid . vsako leto tudi mnogi zbolijo in tudi lahko
2020-07-13	KustecLpicerSimona	nacionalnih športnih federacij. ¶ In zdaj na covid . vprašanje - Kaj smo naredili s tremi
2020-07-13	KustecLpicerSimona	zaposleni preko ostalih oblik, ki jih je covid . nekako zaobel, kot pomoč v času, ko nismo mogli
2020-07-13	KovačičAljaž	upoštevati nekih novih oblik, ki so prišle s tem covidom . me zanima tudi: ¶ Kakšna je vaša ocena
2020-07-13	KustecLpicerSimona	covidsituacije in na delovanja družtev. V času covid . niče ni deloval, tudi običajnih stroškov
2020-07-13	Prebilnik	evropsko kohezijsko politiko in njen razvoj po covidu oziroma med covidom; nekako o temi, ki je za
2020-07-13	Prebilnik	politiko in njen razvoj po covidu oziroma med covidom . nekako o temi, ki je za nekake države, ki
2020-07-13	ČrnaZvonko	tematskih ciljev je bil pogojen z ukrepi covid . in v tem delu je tudi odstopanje mogoče,
2020-07-13	ČrnaZvonko	, da se na posameznih oseh ne glede na situacijo covid . pojavljajo problemi, so se pojavljali že do
2020-07-13	ČrnaZvonko	odstopanje zaradi prerazporeditev za potrebe covid . , tako kot določajo veljavne uredbe iz marca in
2020-07-10	StarovicVojko	nabave mask in respiratorjev med epidemijo covid . . Kako ironično, da je ustanoviteljica
2020-05-26	TomicVioleta	bi poslanci od njih pričakovali. V redu, zaradi covid . smo se izogibali [§§:insaudible: [...]] Počakajte, da končam, ne
2020-05-30	TomicVioleta	izjemno nesposobnost. Žal, če ne bi bilo krize s covidom . , bi si jo ta vlada morala zmisлити, kajti njene
2020-05-14	TomicVioleta	. Za božjo voljo želeli smo sodelovati že pri /MISC: covid . 1 / in COVID 2 zakonu pa smo videli kako ste nas
2020-05-14	TomicVioleta	voljo želeli smo sodelovati že pri /MISC: COVID 1 / in covid . 2 zakonu pa smo videli kako ste nas zaobli in
2020-05-14	ToninMatej	opreme zagotovili, potem nebi mogli zdraviti covid . bolnikov, ker so dejali, da nimajo srca, da bi v

ParlaMint-GALICIA: estado actual

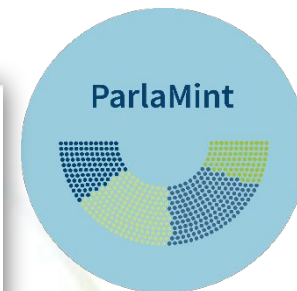
- **Participantes: ILG (USC)**

- Elisa Fernández Rei
- Adina Vladu
- Carmen Magariños
- Daniel Bardanca

- **Datos:**

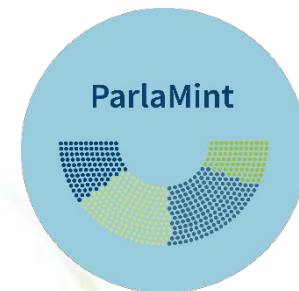
- **Datos fuente:** 277 sesiones plenarias (2015-2021)
- **Metadatos:** 220 interventores, partidos políticos, GPs, legislaturas y gobiernos.

- Está completado el procesado de datos.



ParlaMint-GALICIA: trabajo futuro

- Prueba y validación de los *scripts* de conversión
- Prueba y evaluación de las herramientas de anotación para el gallego
- ParlaSpeech-GA: Creación de un corpus alineado con el audio para el reconocimiento del habla (ASR). Se espera que esté disponible a finales de 2022



ParlaMint-CAT: estado actual

ParlaMint

- **Participantes: Grupo ILATERM-TRL (UPF)**
Marilina Pisani, Rodolfo Zevallos, Núria Bel

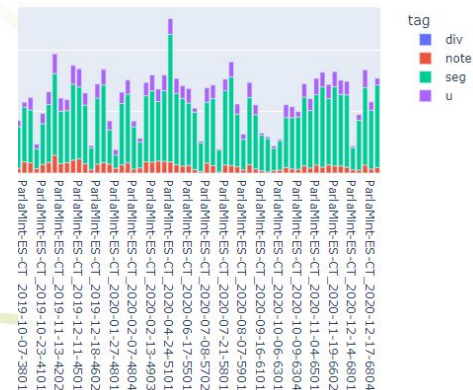
- **Datos procesados:**

211 documentos (2015-2020) por Parlament de Catalunya

Discursos totales: 39.415
Párrafos total: 184.606
Palabras total: 11.680.879
Interrupciones etiquetadas: 15.945
Comentarios etiquetados: 12.354

- **Procesado de datos:**

- Programa para Docx2TEI(xml) + reutilizar comentarios



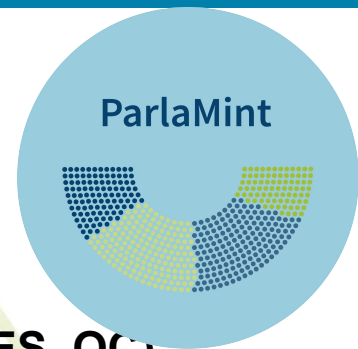
28 299

COMENTARIOS
DEL TRANSCRIPTOR

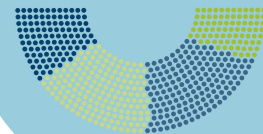
255

INTERVINIENTES
ÚNICOS

ParlaMint-CAT: estado actual



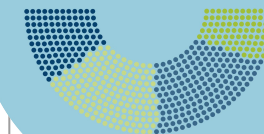
- **Selección de Freeling para anotar lingüísticamente textos (CA, ES, OC)**
- **Anotación lingüística:**
 - Segmentación y tokenización
 - Lematización y categorías gramaticales
 - UD Dependencias sintácticas
 - Marcas Nombres Propios (NERC): PER, LOC, ORG, MISC



- **Participantes:** Mikel Iruskietta, Jon Alkorta, Kike Fernandez.
- **Obtención de datos:**
 - Datos fuente:
 - De 02/2015 a 02/2022.
 - 313 sesiones parlamentarias.
 - Metadatos: datos de 176 interventores, 9 partidos políticos, Parlamento Vasco y Gobierno Vasco.
- **Estructuración de datos:**
 - Dos versiones de un mismo corpus:
 - Corpus original (en euskera y castellano)
 - 1,98 (eus) + 7,35 (cas) + 0,05: 9,38 millones de palabras.
 - Corpus en euskera (con partes originales y traducidas)
 - 7,37 millones de palabras.

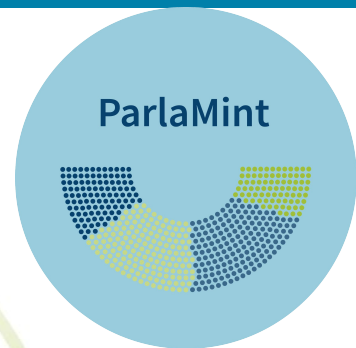
ParlaMint-EUS

ParlaMint



Tarea	Corpus original	Corpus en euskera
Documento de metadatos en formato XML-TEI	+	+
Sesiones parlamentarias en formato XML-TEI	+	+
Identificación del idioma	+	
Identificar a los interventores y su rol	+	X
Identificar el principio y final de cada intervención	+	X
Segmentar las intervenciones en oraciones	+	X
Anotación lingüística: <ul style="list-style-type: none">● Segmentación y tokenización● Lematización y categorías gramaticales.● <i>Universal Dependencies</i>● <i>Named Entities Recognition</i>	X	X

Para más información



- Página web:
 - <https://www.clarin.eu/parlamint>.
- Artículo de referencia sobre el proyecto:
 - Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. [*The ParlaMint corpora of parliamentary proceedings*](#). *Language Resources and Evaluation*, 2022.