ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions

Baybars Külebi, Carme Armentano-Oller, Carlos Rodríguez-Penagos, Marta Villegas
Outline

• The motivation
• Background
• Technical implementation
• Results
• Future
The motivation

The digital gap is recognized by the European Parliament

P8_TA(2018)0332

Language equality in the digital age

European Parliament resolution of 11 September 2018 on language equality in the
digital age (2018/2028(INI))

Current obstacles to achieving language equality in the digital age in Europe

1. Regrets the fact that, owing to a lack of adequate policies in Europe, there is currently a
widening technology gap between well-resourced languages and less-resourced
languages, whether the latter are official, co-official or non-official in the EU; regrets,
furthermore, the fact that more than 20 European languages are in danger of digital
language extinction; notes that the EU and its institutions have a duty to enhance,
promote and uphold linguistic diversity in Europe;
The motivation

According to recent surveys, the youth in Catalonia is speaking less and less Catalan.

Current obstacles to achieving language equality in the digital age in Europe

1. Regrets the fact that, owing to a lack of adequate policies in Europe, there is currently a widening technology gap between well-resourced languages and less-resourced languages, whether the latter are official, co-official or non-official in the EU; regrets, furthermore, the fact that more than 20 European languages are in danger of digital language extinction; notes that the EU and its institutions have a duty to enhance, promote and uphold linguistic diversity in Europe;
14/02/2020
Conference on virtual assistants and Catalan

10/12/2020
Presentation of AINA

Neix ‘AINA’, el projecte del Govern per garantir el català en l’era digital
Technical outline
Consell de Catalunya, 17 de gener de 2018

Sessió núm. 1

La sessió s'obre a les once del matí i dos minuts. Presideix el president de la Mesa d'Elecció, acompanyat dels secretaris de la Mesa d'Elecció, el qual és assistit per el secretari general i el fetat major.

ORDRE DEL DIA DE LA CONVOCATÒRIA
Puntunic: Constitució del Ple del Parlament i elecció de la Mesa del Parlament ( tram. 396-000012/18 i 396-000013/18).

El secretari general (Xavier Muro i Bass)

Bo diu a tot hom, illustres senyores diputats i senyores diputades, autoritzats que ens acompanyen, senyores i senyores, signen-hem benvinguts al Parlament de Catalunya en aquesta sessió constitutiva de la seua dotzena legislatura.

Constitució del Ple del Parlament i elecció de la Mesa del Parlament

396-000012/18 i 396-000013/18

Tal com és preceptiu, s'ha d'obrir amb la lectura del Decret de convocatòria, que és el Reial decret 1013/2018, de 6 de gener, de convocatòria de la sessió constitutiva del Parlament de Catalunya, publicat al Boletí Oficial de l'Estat núm. 9, de 10 de gener, i al Diari Oficial de la Generalitat de Catalunya, núm. 732/A, de 10 de gener.

Aquest decret diu així: «Realitzades les eleccions al Parlament de Catalunya el passat dia 21 de desembre i atès que les corresponents joves electors provinents han proclamat els resultats d'aquest procés electoral.

Vist el que disposa l'article 10.2, de la Llei 13/2008, del 5 de novembre, de la presidència de la Generalitat i del Govern.

De conformitat amb el que estableixen els paràgrafs segons i tercer de l'apartat A de l'acord del Consell de Ministres de 21 d'octubre de 2017, publicat mitjançant l'Ordre de presidència 1034/2017, de 27 d'octubre, en el Boletí Oficial de l'Estat núm. 206, de 27 d'octubre, així com amb l'article 3 del Reial decret 944/2017, de 27 d'octubre, pel qual es designen òrgans i autoritzats encarregats de donar compliment a les mesures dirigides al Govern i a l'Administració de la Generalitat de Catalunya, autoritzades per l'acord del Ple del Senat, de 27 d'octubre de 2017, pel qual s'aprovaren les mesures requerides pel Government, a l'espai de l'article 155 de la Constitució, decreta.

...
With the support of Departament de Cultura and the AINA project we generated **611 hours** of speech corpus.

- Segmented
- Speaker id
- Gender
- w/o punctuation
Extraction Flux
Web parser
El president

Començant la sessió. La llista de preguntes a respondre oralment en el Ple està inclosa en el dossier que tenen tots vots, i, d'acord amb l'article 164 del Reglament, se substanciaran demà al matí, a les deu del matí, tal com apareix al setè punt de l'ordre del dia d'aquesta sessió plenària.

Comunicació al Ple de la composició dels mesos de les comissions i de la Diputació Permanent (arts. 49.2 i 74.3 del Reglament)

El primer punt de l'ordre del dia és la comunicació al Ple de la composició de les mesos de les comissions i Diputació Permanent. D'acord amb això que estableix els articles 49.2 i 74.3 del Reglament, la composició de les mesos de les comissions i de la Mesa de la Diputació Permanent, ha de ser comunicada al Ple.

Atès que la nova composició de les mesos de les comissions i de la Diputació Permanent està inclosa al dossier del Ple que tenen tots vots, els prego que se n'executin la lectura. SI7 (Pausa.) Per tant, poden donar així per completar això que estableix el Reglament i, per tant, he informat al Ple de la composició d'aquestes mesos. Molt bé.

Interpel·lació al Govern sobre les polítiques de suport a l'empresa

S00-00018/12

Doncs passem al segon punt de l'ordre del dia, que són les interpel·lacions; i, en aquest cas, la interpel·lació al Govern sobre les polítiques de suport a l'empresa, que presenta el Grup Parlamentari Socialistes i Units per Avançar. Exposa la interpel·lació la diputada Alicia Romero.

Alicia Romero Llanas

Bona tarda, diputats i diputades. Consellers... bé, avui el Grup Socialista i Units per Avançar vol parlar sobre... d'alguna manera sobre la ambigües d'un país, no?, sobre quines són les estratègies de Catalunya pel que fa, doncs, a la generació de la seva riquesa. I ho dig així, en general, perquè creiem que és el que ens agradaria fer. Com a mínim, ens agradaria fer una reflexió de país per construir i per serrar. Per
El presidente: Comencemos la sesión. La lista de preguntas a responder oralmente en el Pleno está incluida en el dossier que tienen todos voto. Si, d.u.2019, acordó con la u.2019 parte 164 del Reglamento, se substanciaron deniek el mat. (X) y el d.u.2019 parte 164 del Reglamento, se substanciaron deniek el mat. (X).

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.

El presidente: La composición de las mesas de las comisiones y de la Diputación Permanente (artículos 49.2 y 74.3 del Reglamento). El primer punto de la u.2019 el diputado el (X) se substanciaron deniek el mat. (X) a la composición de las mesas de las comisiones y de la Diputación Permanente. D.u.2019, acordó con el (X) que estableció el artículo 49.2 y 74.3 del Reglamento.
Alignment

Problem:

Web: H. Sr. Oriol Junqueras i Vies (Conseller)
Pdf: El vicepresident del Govern i conseller d’Economia i Hisenda (Oriol Junqueras i Vies)

Web: H. Sr. Lluís Miquel Recoder i Miralles (Conseller)
Pdf: El conseller de Territori i Sostenibilitat

Solution

Initial text matching
Oriol Junqueras i Vies is common in both

• Sequence alignment using Smith-Waterman algorithm

\[
\begin{align*}
A & \ C & T & A & C \\
A & \ C & G & A & C \\
\end{align*}
\]

\[T = G\]
Text alignment and Segmentation

Long text alignment is undertaken using a basic ASR model, finite state graphs using the transcribed text (Panoyotov et al. 2015)

Segmentation takes advantage of silences and punctuation, via a "modeled beam search"

All segments are scored
Results

| Subcorpus     | Gender | Duration (h) |
|---------------|--------|--------------|
| other_test    | F      | 2.516        |
| other_dev     | F      | 2.701        |
| other_train   | F      | 109.68       |
| other_test    | M      | 2.631        |
| other_dev     | M      | 2.513        |
| other_train   | M      | 280.196      |
| **other total** |       | **400.239**  |
| clean_test    | F      | 2.707        |
| clean_dev     | F      | 2.576        |
| clean_train   | F      | 77.905       |
| clean_test    | M      | 2.516        |
| clean_dev     | M      | 2.614        |
| clean_train   | M      | 123.162      |
| **clean total** |       | **211.48**   |
| **Total**     |        | **611.719**  |

Table 1: The total duration of all the subsets, with gender distribution.

The corpus is divided into clean and other, according to the segment scores.

Segments are tagged with the gender of the speaker, according to their titles (Sr., Sra.)

dev, train, val subcorpora do not have overlapping speakers

Data in Zenodo and Hugging face
Future work

- Training ASR models with and without ParlamentParla
- Process the content, post 2018/07
- Process the content of commissions
- IT department of Parliament is going to provide an API for easy access to the data.
- Automate generation of speech corpus via API for continuous update of the speech corpus
Thank you