

ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

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

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;

LLENGUA

 **Només un de cada tres joves de Barcelona parla català de manera habitual**

- La meitat dels que tenen entre 15 i 34 anys consideren que tenen un bon nivell de llengua

Actualitzada el 18-08-2021 19:28

Un 28% dels joves que viuen a Barcelona parlen català de manera habitual. Així ho diu l'**Enquesta de Joventut** publicada per l'Ajuntament de la capital catalana. Això suposa **una davallada del 7% en comparació amb l'enquesta del 2015**, quan eren un 35% els joves que feien servir el català com a llengua habitual. Tot i no fer-lo servir, **més de la meitat dels enquestats reconeixen que tenen un bon domini del català.**

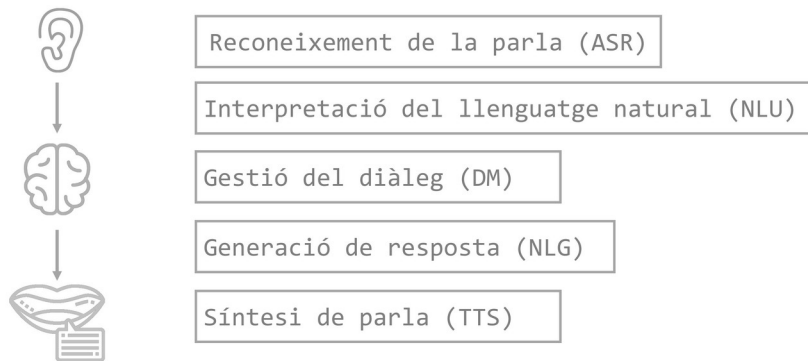
Background

Jornada sobre assistents de veu i llengua catalana

#CatalàDigital

Infraestructura pública d'assistents de veu

L'ecosistema



 Generalitat de Catalunya

 movistar CENTRE

 fundació .cat

14/02/2020

Conference on virtual assistants and Catalan

10/12/2020
Presentaion of AINA

Neix 'AINA', el projecte del Govern per garantir el català en l'era digital

BY RECEPCIÓN COMUNICADOS · 10 DE DECEMBER DE 2020



Technical outline

Resource parlament.cat

» Seqüència del debat



Sobre la sessió.

Ordenació del debat.

Intervinent: Sr. Xavier Muro i Bas (Secretari General)

De 11:02:09 a 11:02:25 - Durada: 0 m. 16 s.

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Lectura.

▣ Punts tractats conjuntament

Intervinent: Sr. Xavier Muro i Bas (Secretari General)

De 11:02:25 a 11:06:31 - Durada: 4 m. 5 s.

Diari de sessions: **DSPC-P 001/12**

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Intervenció.

▣ Punts tractats conjuntament

Intervinent: H. Sr. Ernest Maragall i Mira

De 11:06:31 a 11:17:47 - Durada: 11 m. 16 s.

Diari de sessions: DSPC-P 001/12

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Lectura.

▣ Punts tractats conjuntament

Intervinent: I. Sra. Rut Ribas i Martí

De 11:17:47 a 11:23:13 - Durada: 5 m. 25 s.

Diari de sessions: DSPC-P 001/12

» Ple del Parlament. 17/01/2018 - sessió constitutiva

DSPC-P 1
17 de gener de 2018

SESSIÓ NÚM. 1

La sessió s'obre a les onze del matí i dos minuts. Presideix el president de la Mesa d'Edat, acompanyat dels secretaris de la Mesa d'Edat, la qual és assistida pel secretari general i el lletrat major.

ORDRE DEL DIA DE LA CONVOCATÒRIA

Punt únic: Constitució del Ple del Parlament i elecció de la Mesa del Parlament (tram. 396-00001/12 i 398-00001/12).

El secretari general (Xavier Muro i Bas)

Bon dia a tothom, il·lustres senyores diputades i senyores diputats, autoritats que ens acompanyen, senyores i senyors, sigueu benvinguts al Parlament de Catalunya en aquesta sessió constitutiva de la seva dotzena legislatura.

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

396-00001/12 i 398-00001/12

Tal com és preceptiu, d'acord amb l'article 1 del Reglament de la cambra, aquesta sessió s'ha d'iniciar amb la lectura del Decret de convocatòria, que és el Reial decret 1/2018, de 9 de gener, de convocatòria de la sessió constitutiva del Parlament de Catalunya, publicat al *Bulletí Oficial de l'Estat* número 9, de 10 de gener, i al *Diari Oficial de la Generalitat de Catalunya*, número 7532A, de 10 de gener.

Aquest decret diu així: «Realitzades les eleccions al Parlament de Catalunya el propassat dia 21 de desembre i atès que les corresponents juntes electorals provincials han proclamat els resultats d'aquest procés electoral.

»Vist el que disposa l'article 10.d, 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 segon 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 *Bulletí Oficial de l'Estat* número 260, 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 autoritats 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'aproven les mesures requerides pel Govern, a l'empars de l'article 155 de la Constitució, decreto:

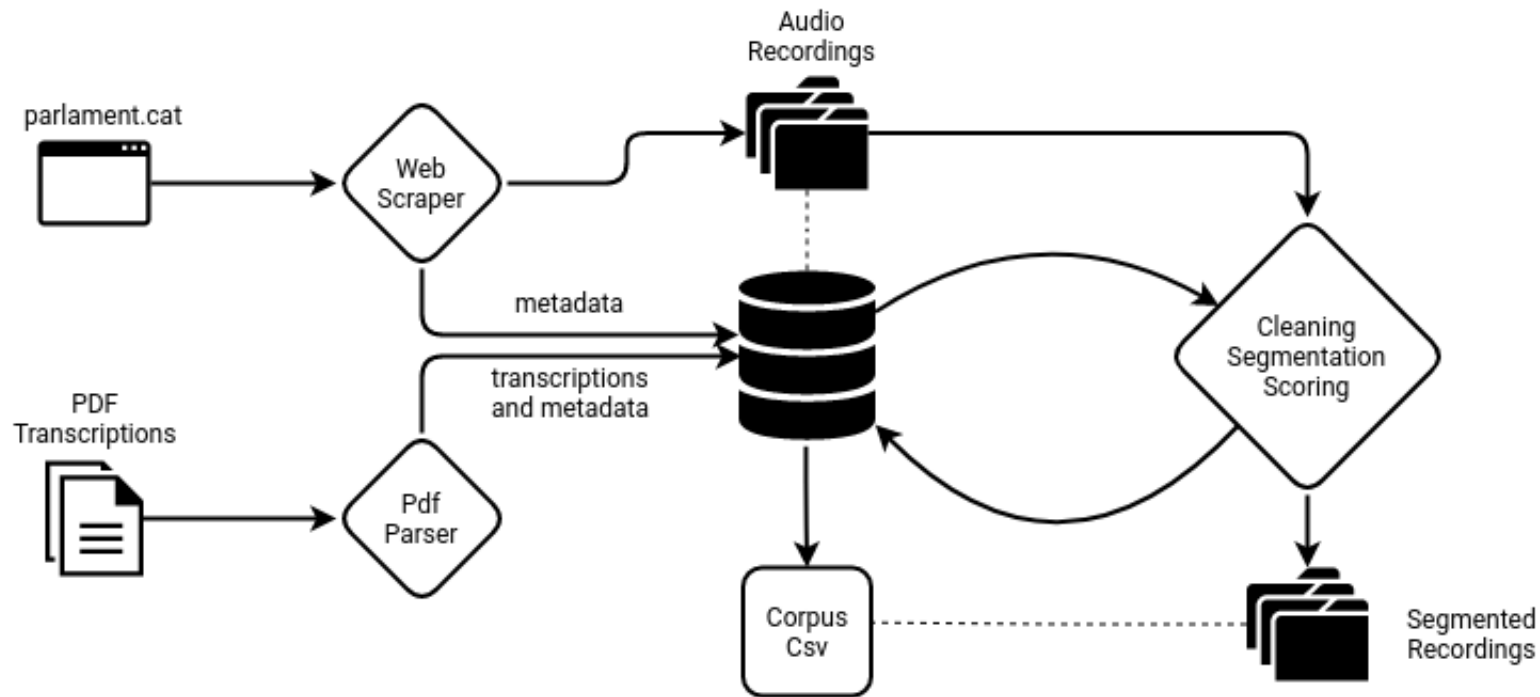
Resource parlament.cat

speaker_id	path	sentence	gender	duration
64	clean_dev/e/e/ee5d4bc297152a3f3e21_26.1_33.52.wav	el primer punt aquest punt el votarem a favor perquè ja el nostre grup parlamentari va presentar unes propostes de resolució	F	7.42
64	clean_dev/e/e/ee5d4bc297152a3f3e21_34.08_48.88.wav	l'any passat que anaven en la línia de desenvolupar un pla d'estalvi i eficiència energètica per a catalunya per tal de definir l'estratègia en tres sectors importants que són el de l'edificació el transport i la indústria per tant donarem suport a aquest primer punt	F	14.8
64	clean_dev/e/e/ee5d4bc297152a3f3e21_73.97_79.38.wav	per tant també votarem a favor d'aquest punt pel que fa al punt tercer	F	5.41
64	clean_dev/e/e/ee5d4bc297152a3f3e21_79.95_93.71.wav	estem d'acord que cal preparar una cimera de la convenció catalana del canvi climàtic perquè això suposarà poder desenvolupar objectius i unir estratègies que ens portaran a avançar cap al dos mil dotze i el dos mil vint	F	13.76
64	clean_dev/e/e/ee5d4bc297152a3f3e21_120.09_132.14.wav	perquè això comporta una pressió econòmica i tenint en compte doncs que hi ha hagut una pujada d'impostos recentment creiem que no és el moment de penalitzar ciutadans ni empreses que estan patint els efectes greus	F	12.05
64	clean_dev/e/e/ee5d4bc297152a3f3e21_186.44_191.61.wav	el preàmbul de la llei justifica fer una parada temporal d'aquestes ajudes a les renovables	F	5.17
64	clean_dev/e/e/ee5d4bc297152a3f3e21_191.74_204.72.wav	eis últims anys el creixement de les tecnologies incloses en el règim especial ha permès superar en el dos mil deu els objectius de potència instal·lats previstos en el pla d'energies renovables dos mil cinc dos mil deu	F	12.98
64	clean_dev/e/e/ee5d4bc297152a3f3e21_206.45_220.37.wav	per tant actualment a espanya hi ha una capacitat de producció d'energia instal·lada suficient per cobrir la demanda prevista per als propers anys per tant no és necessari continuar instal·lant potència de forma que a mitja	F	13.92
326	clean_dev/f/8/f885d8e8834c09f9bdf2_69.82_78.06.wav	el diputat parlarà una deducció del tres per cent de la quota íntegra de l'irpf de les quantitats percebudes pel contribuïent	F	8.24
326	clean_dev/f/8/f885d8e8834c09f9bdf2_85.91_100.0.wav	agroambientals de benestar animal en ramaderia ecològica d'inversions no productives en xarxa natura dos mil i de zones de muntanya i d'indemnització compensatòria en zones desfavorides fora de la muntanya	F	14.09
326	clean_dev/f/8/f885d8e8834c09f9bdf2_130.99_137.29.wav	econòmics necessaris de les partides pressupostàries del departament d'agricultura alimentació i acció rural	F	6.3
326	clean_dev/f/8/f885d8e8834c09f9bdf2_153.26_158.88.wav	just al límit entre l'euro atlàntica i la mediterrània alhora que també és un país densament poblat	F	5.62
326	clean_dev/f/8/f885d8e8834c09f9bdf2_160.03_170.95.wav	la salvaguarda d'aquesta diversitat biològica sotmesa a una forta pressió del seu entorn social exigeix l'aplicació de mesures i instruments de gestió avançats i eficaços	F	10.92
326	clean_dev/f/8/f885d8e8834c09f9bdf2_171.39_179.52.wav	que despleguin les directrius que emanen de la legislació europea com és l'aplicació de la xarxa natura dos mil a catalunya	F	8.13
326	clean_dev/f/8/f885d8e8834c09f9bdf2_206.15_212.17.wav	i estèpic s'incrementa més d'un trenta-cinc per cent l'àmbit de muntanya mediterrània doble	F	6.02
326	clean_dev/f/8/f885d8e8834c09f9bdf2_212.39_222.73.wav	les hectàrees protegides i l'àmbit fluvial arriba gairebé al noranta per cent fet que permetrà el manteniment dels boscos de ribera i de la fauna que aquests acullen	F	10.34
326	clean_dev/f/8/f885d8e8834c09f9bdf2_223.03_231.37.wav	però paral·lelament a la proposta de redefinició dels límits de la xarxa natura dos mil s'està treballant per dotar	F	8.34

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

» Seqüència del debat



Sobre la sessió.

Ordenació del debat.

Intervinent: Sr. Xavier Muro i Bas (Secretari General)

De 11:02:09 a 11:02:25 - Durada: 0 m. 16 s.

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Lectura.

Punts tractats conjuntament

Intervinent: Sr. Xavier Muro i Bas (Secretari General)

De 11:02:25 a 11:06:31 - Durada: 4 m. 5 s.

Diari de sessions: DSPC-P 001/12

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Intervenció.

Punts tractats conjuntament

Intervinent: H. Sr. Ernest Maragall i Mira

De 11:06:31 a 11:17:47 - Durada: 11 m. 16 s.

Diari de sessions: DSPC-P 001/12

» Ple del Parlament. 17/01/2018 - sessió constitutiva



Lectura.

Punts tractats conjuntament

Intervinent: I. Sra. Rut Ribas i Martí

De 11:17:47 a 11:23:13 - Durada: 5 m. 25 s.

Diari de sessions: DSPC-P 001/12

» Ple del Parlament. 17/01/2018 - sessió constitutiva



```
{
  "_id" : "385d0964823380132dcbea2eaafd1638",
  "value" : {
    "ple_code" : "2017_05_18_217559",
    "urls" : [
      [
        "H. Sra. Meritxell Ruiz i Isern (Consellera) "
      ],
      "/home/baybars/repositories/parlament-scrape/audio/
8/4/8411b4c23389529a41c0.mp3"
    ]
  ]
}
{
  "_id" : "0e4b3a9413c1bc9183f34ddc12ba65b2",
  "value" : {
    "ple_code" : "2017_05_18_217559",
    "urls" : [
      [
        "Sra. Eulàlia Reguant i Cura (Membre) "
      ],
      "/home/baybars/repositories/parlament-scrape/audio/e/d/
edee3c4519054ca4c861.mp3"
    ]
  ]
}
```

presente

Pdf parser

DSPC-P 17
17 de juliol de 2018

El president

Comencem la sessió. La llista de preguntes a respondre oralment en el Ple està inclosa en el dossier que tenen tots vostès, 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 de les meses 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 meses de les comissions i Diputació Permanent. D'acord amb allò que estableixen els articles 49.2 i 74.3 del Reglament, la composició de les meses de les comissions i de la Mesa de la Diputació Permanent, ha de ser comunicada al Ple.

Atès que la nova composició de les meses de les comissions i de la Diputació Permanent està inclosa al dossier del Ple que tenen tots vostès, els prego que se n'eximeixi la lectura. Sí? *(Pausa.)* Per tant, podem donar així per complet allò que estableix el Reglament i, per tant, he informat al Ple de la composició d'aquestes meses. Molt bé.

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

300-00018/12

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 Àlicia Romero.

Àlicia Romero Llano

Bona tarda, diputats i diputades. Conselleres..., bé, avui el Grup Socialistes i Units per Avançar vol parlar sobre..., d'alguna manera, sobre les ambicions d'un país, no?, sobre quines són les estratègies de Catalunya pel que fa, doncs, a la generació de la seva riquesa. I ho dic 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 sumar. Per

```
<page number="4" position="absolute" top="0" left="0" height="1262" width="892">
<fontspec id="11" size="14" family="Times" color="#231f20"/>
<fontspec id="12" size="14" family="Times" color="#231f20"/>
<text top="90" left="85" width="61" height="13" font="2">DSPC-P 17</text>
<text top="105" left="85" width="106" height="13" font="2">17 de juliol de 2018</text>
<text top="1196" left="85" width="60" height="13" font="2">Sessió 14</text>
<text top="1196" left="802" width="6" height="11" font="2">1</text>
<text top="86" left="296" width="82" height="16" font="10">El president</text>
<text top="107" left="296" width="517" height="16" font="11">Comencem la sessió. La llista de preguntes
a respondre oralment en el Ple està
<text top="127" left="274" width="537" height="18" font="11">inclosa en el dossier que tenen tots
vostès, i, d'acord amb l'article 164 del Reglament,
<text top="147" left="274" width="538" height="18" font="11">se substanciaran demà al matí, a les deu
del matí, tal com apareix al setè punt de
<text top="168" left="274" width="252" height="18" font="11">l'ordre del dia d'aquesta sessió plenària.
</text>
<text top="229" left="296" width="496" height="17" font="0">Comunicació al Ple de la composició de les
meses de les comissions i
<text top="248" left="296" width="409" height="17" font="0">de la Diputació Permanent (arts. 49.2 i
74.3 del Reglament)</text>
<text top="279" left="296" width="516" height="18" font="11">El primer punt de l'ordre del dia és la
comunicació al Ple de la composició de les
<text top="299" left="274" width="538" height="18" font="11">meses de les comissions i Diputació
Permanent. D'acord amb allò que estableixen
<text top="319" left="274" width="538" height="18" font="11">els articles 49.2 i 74.3 del Reglament, la
composició de les meses de les comissions i
<text top="340" left="274" width="431" height="18" font="11">de la Mesa de la Diputació Permanent, ha
de ser comunicada al Ple.</text>
<text top="360" left="296" width="518" height="18" font="11">Atès que la nova composició de les meses
de les comissions i de la Diputació
<text top="380" left="274" width="539" height="18" font="11">Permanent està inclosa al dossier del Ple
que tenen tots vostès, els prego que se
<text top="400" left="274" width="538" height="18" font="11">
n'eximeixi la lectura. Sí?
<i>(Pausa.)</i>
Per tant, podem donar així per complet allò que
</text>
<text top="421" left="274" width="539" height="18" font="11">estableix el Reglament i, per tant, he
informat al Ple de la composició d'aquestes
<text top="441" left="274" width="104" height="18" font="11">meses. Molt bé. </text>
<text top="503" left="296" width="456" height="17" font="0">Interpel·lació al Govern sobre les
polítiques de suport a l'empresa</text>
<text top="524" left="296" width="82" height="15" font="1">300-00018/12</text>
<text top="552" left="296" width="517" height="18" font="11">Doncs passem al segon punt de l'ordre del
dia, que són les interpel·lacions; i, en
<text top="573" left="274" width="538" height="18" font="11">aquest cas, la interpel·lació al Govern
sobre les polítiques de suport a l'empresa, que
<text top="593" left="274" width="534" height="18" font="11">presenta el Grup Parlamentari Socialistes
i Units per Avançar. Exposa la interpel·la
<text top="613" left="274" width="195" height="18" font="11">ció la diputada Àlicia Romero Llano.
<text top="643" left="296" width="138" height="16" font="10">Àlicia Romero Llano</text>
<text top="664" left="296" width="516" height="18" font="11">Bona tarda, diputats i diputades.
Conselleres..., bé, avui el Grup Socialistes i Units
<text top="684" left="274" width="538" height="18" font="11">per Avançar vol parlar sobre..., d'alguna
manera, sobre les ambicions d'un país, no?, </text>
```

Pdf parser

```
<?xml version="1.0" encoding="utf-8" ?>
<page number="4" position="absolute" top="0" left="0" height="1262" width="892">
  <fontspec id="11" size="14" family="Times" color="#231f20"/>
  <fontspec id="12" size="14" family="Times" color="#231f20"/>
  <text top="90" left="85" width="61" height="13" font="2">DSPC-P 17</text>
  <text top="105" left="85" width="106" height="13" font="2">17 de juliol de 2018</text>
  <text top="1196" left="85" width="60" height="13" font="2">Sessió 14</text>
  <text top="1196" left="802" width="6" height="13" font="2"></text>
  <text top="86" left="296" width="82" height="16" font="10">El president</text>
  <text top="107" left="296" width="517" height="16" font="11">Comencem la sessió. La llista de preguntes a respondre oralment en el Ple està
  <text top="127" left="274" width="537" height="18" font="11">inclosa en el dossier que tenen tots vostès, i, d'acord amb l'article 164 del Reglament,
  <text top="147" left="274" width="538" height="18" font="11">se substanciaran demà al matí, a les deu del matí, tal com apareix al setè punt de
  <text top="168" left="274" width="252" height="18" font="11">l'ordre del dia d'aquesta sessió plenària.
  <text top="229" left="296" width="496" height="17" font="0">Comunicació al Ple de la composició de les meses de les comissions i
  <text top="248" left="296" width="409" height="17" font="0">de la Diputació Permanent (arts. 49.2 i 74.3 del Reglament)</text>
  <text top="279" left="296" width="516" height="18" font="11">El primer punt de l'ordre del dia és la comunicació al Ple de la composició de les
  <text top="299" left="274" width="538" height="18" font="11">meses de les comissions i Diputació Permanent. D'acord amb allò que estableixen
  <text top="319" left="274" width="538" height="18" font="11">els articles 49.2 i 74.3 del Reglament, la composició de les meses de les comissions i
  <text top="340" left="274" width="431" height="18" font="11">de la Mesa de la Diputació Permanent, ha de ser comunicada al Ple.
  <text top="360" left="296" width="518" height="18" font="11">Atès que la nova composició de les meses de les comissions i de la Diputació
  <text top="380" left="274" width="539" height="18" font="11">Permanent està inclosa al dossier del Ple que tenen tots vostès, els prego que se
  <text top="400" left="274" width="538" height="18" font="11">n'eximeixi la lectura. Sí?
  <i>(Pausa.)</i>
  Per tant, podem donar així per complet allò que
  <text top="421" left="274" width="539" height="18" font="11">estableix el Reglament i, per tant, he informat al Ple de la composició d'aquestes
  <text top="441" left="274" width="104" height="18" font="11">meses. Molt bé.
  <text top="503" left="296" width="456" height="17" font="0">Interpel·lació al Govern sobre les polítiques de suport a l'empresa
  <text top="524" left="296" width="82" height="15" font="1">300-00018/12</text>
  <text top="552" left="296" width="517" height="18" font="11">Doncs passem al segon punt de l'ordre del dia, que són les interpel·lacions; i, en
  <text top="573" left="274" width="538" height="18" font="11">aquest cas, la interpel·lació al Govern sobre les polítiques de suport a l'empresa, que
  <text top="593" left="274" width="534" height="18" font="11">presenta el Grup Parlamentari Socialistes i Units per Avançar. Exposa la interpel·lació
  <text top="613" left="274" width="195" height="18" font="11">de la diputada Alicia Romero.
  <text top="643" left="296" width="138" height="16" font="10">Alicia Romero Llanó</text>
  <text top="664" left="296" width="516" height="18" font="11">Bona tarda, diputats i diputades. Conselleres... bé, avui el Grup Parlamentari Socialistes i Units
  <text top="684" left="274" width="538" height="18" font="11">per Avançar vol parlar sobre..., d'alguna manera, sobre les ambicions d'un país, no?,</text>
</page>
```



```
- {El president: "Comencem la sessió. La llista de preguntes a respondre oralment en el Ple està inclosa en el dossier que tenen tots vostès, 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 de les meses 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 meses de les comissions i Diputació Permanent. D'acord amb allò que estableixen els articles 49.2 i 74.3 del Reglament, la composició de les meses de les comissions i de la Diputació Permanent, ha de ser comunicada al Ple. Atès que la nova composició de les meses de les comissions i de la Diputació Permanent està inclosa al dossier del Ple que tenen tots vostès, els prego que se n'eximeixi la lectura. Sí? (Pausa.) Per tant, podem donar així per complet allò 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 300-00018/12 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 de la diputada Alicia Romero."}
- {"Alícia Romero Llanó": "Bona tarda, diputats i diputades. Conselleres..., avui el Grup Parlamentari Socialistes i Units per Avançar vol parlar sobre..., 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 de les meses 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 meses de les comissions i Diputació Permanent. D'acord amb allò que estableixen els articles 49.2 i 74.3 del Reglament, la composició de les meses de les comissions i de la Diputació Permanent, ha de ser comunicada al Ple. Atès que la nova composició de les meses de les comissions i de la Diputació Permanent està inclosa al dossier del Ple que tenen tots vostès, els prego que se n'eximeixi la lectura. Sí? (Pausa.) Per tant, podem donar així per complet allò 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 300-00018/12 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 de la diputada Alicia Romero."}
```

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

A C T A C

A C G A C

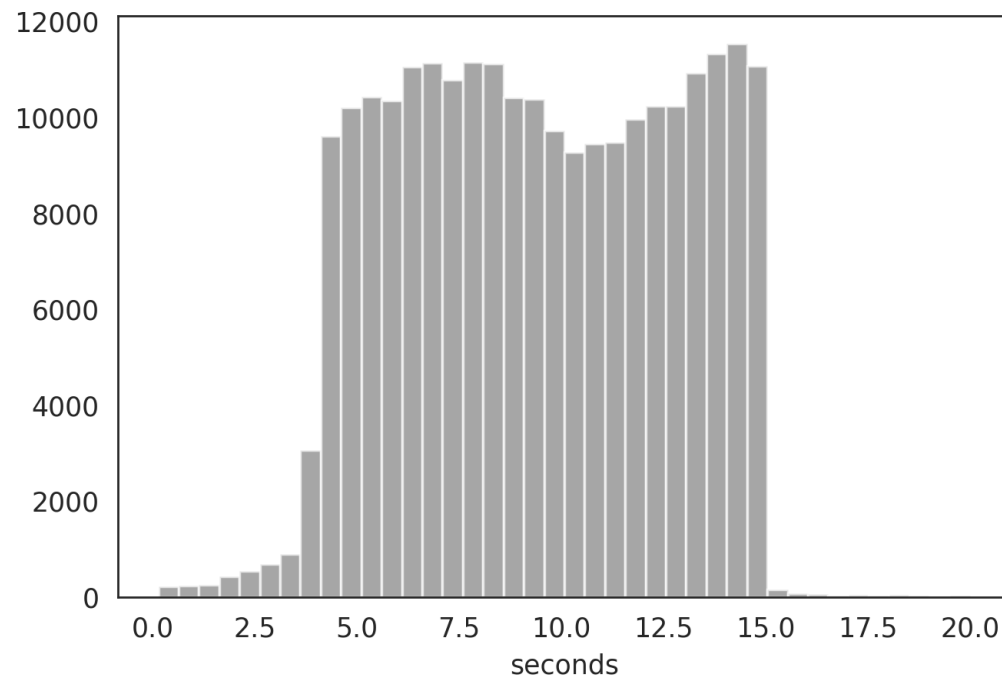
→ T = G

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 punctuations, 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 ParliamentParla
- 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