

Corpus d'évaluation du
Projet ETAPE

Matthieu Carré, Niklas Paulsson
{carre,paulsson}@elda.org

ELDA

Plan

- Introduction
- Emissions
- Collecte & Sélection
- Transcription
- Annotation en Entités Nommées (EN)
- Contrôle qualité
- Conclusion

Emissions

Emissions (1/2)

- **BFM Story:** informations, reportages & débats
- **LCP Pile et face:** débats entre deux invités politiques (+présentateur)
- **LCP Ca vous regarde:** analyses, débats, reportages et témoignages en direct
- **LCP Entre les lignes:** débats d'actualité par 4 chroniqueurs (news magazines) + présentateur
- **LCP Top questions:** questions au gouvernement (assemblée nationale)

Emissions (2/2)

- BFM
 - BFM Story (60min)
- LCP
 - Ça vous regarde (50min)
 - Entre les lignes (30 min)
 - Pile et face (30 min)
 - Top Questions (15 min)
- TV8 Mont Blanc
 - La place du village (30 min)

Collecte & Sélection

Collecte

- Volume: ~158 heures
 - BFM TV, LCP (TNT, capté à Paris)
 - TV 8 Mont Blanc (Freebox)
- Vidéo au format *mpeg-2*
- Extraction/Conversion audio au format *wav* (16 bits, 16 et 48 kHz)

Sélection ~29h

Fichier	Enregist. (min)	Parole (min)	Superpos. (min)
BFM Story	442	332	8
LCP Pile et face	314	264	20
LCP Ca vous regarde	289	249	24
LCP Entre les lignes	347	270	34
LCP Top questions	206	136	1
TV8 La place du village	135	94	3
	1733	1345	90 (6,7%)
	28,9 h	22,4 h	1,5 h

Transcription

Transcription

- *Transcriber* avec la vidéo (wav + avi)
- Conventions Ester 2
- Marquer la parole superposée:
 - Signe de reconnaissance (« Back channel »)
 - Tentative de prise de parole (« Turn stealing »)
 - Superposition sur transition (« Anticipated turn taking »)
 - Complément (« Complementary »)

Transcriber

Applications Raccourcis Système | sam. 20 nov., 16:23 | elda

Transcriber 1.5.2

File Edit Signal Segmentation Options Help

- la bataille du Front National a démarré officiellement mercredi mais la campagne interne de Marine Le Pen démarre aujourd'hui dans le Var à Cuers.
- Marine Le Pen, qui est avec nous, et on la salue dans un instant; le temps pour moi d'accueillir Jean-Yves Camus, [r] le politologue spécialiste du Front National;
- on va bien sûr parler du Front National mais aussi, de l'actualité.

speaker#2

- Marine Le Pen, bonsoir! [b] merci d'être en direct dans

speaker#2 + Marine Le Pen

- 1: BFM Story
- 2: bonsoir

speaker#2

- alors avant de parler euh de la compétition interne qui commence au sein de votre mouvement, [r] question d'actualité!

speaker#2

- êtes-vous choquée par la remise en liberté du second braqueur du casino d'Uriage, second braqueur présumé?

(no speaker)

- [b]

Marine Le Pen

- [b-] mais qui ne le serait pas?
- [r] euh moi je ma première pensée dans les forces de l'ordre [r] parce que depuis déjà de trop nombreux mois, elles sont désarmées moralement, [r] elles sont humiliées et elles sont évidemment aujourd'hui [r] bafouées [r] par cette décision. [-b]
- décision qu'il faut mettre en parallèle avec celle [r] euh qui a maintenue en détention René Galinier, [r]
- ce septuagénaire [r] qui mort de peur a tiré euh [r] ce que d'ailleurs, [r] euh je critique bien sûr [r] mais euh a tiré sur deux cambrioleuses qui avaient.

BFMTV_test.trs
BFMTV_BFMStory_2010-09-03_175900.wav

speaker#2	sp	speaker#2	speaker#2	(no ...	[b]	[b-]	[r]	[r]
Marine Le...	B.	alors avant de parler euh de...	êtes-vous choquée par la...	[b]	[b-]	[r]	[r]	euh moi je ma première pensée dans les force
...direct dans	bo	... d'actualité!	... présumé?					... [r] bafouées [r] par cette décision. [-b]
6:10		6:15	6:20	6:25	6:30	6:35		

Cursor : 06:23.057

bin - Navigateur de fi... | Transcriber 1.5.2 | [gnome-screenshot - ...]

Parole superposée

Applications Raccourcis Système mer. 8 déc., 12:50 elda

Transcriber 1.5.2

File Edit Signal Segmentation Options Help

Hervé Gaymard

- `[dis=dm-]ben[-dis=dm]` Michel Sapin dit à la fin de son propos le contraire de ce qu'i(1) dit au début `[b-]je suis complètement[-b]` d'accord avec lui `[r]` pour dire
- `(/[dis=rep-]qu'on ne peut pas rester qu'on ne peut pas[-dis=rep]/)` `(/sapin: non mais d'accord avec le début mais pas avec la fin [r]/c)`
- qu'on ne peut pas rester avec des niveaux `[dis=hes-]euh[-dis=hes]` de déficit qui nourrissent l'endettement `[dis=hes-]euh[-dis=hes]` et
- dans cette configuration `[dis=hes-]euh[-dis=hes]` que faut-il faire `[dis=dm-]sbah[-dis=dm]` il faut bien évidemment modérer la dépense et toutes les dépenses, pas seulement celles de l'état mais aussi celles des collectivités locales et c'est un `[r]` un président de conseil général qui le dit et qui l'assume `[r]`
- et puis il faut mettre fin aussi `[dis=hes-]euh[-dis=hes]` à un certain nombre d'avantages fiscaux qui n'avaient jamais été réexaminés dans le passé `[r]`
- et *qui étaient reconduits automatiquement donc on a un déficit qui baisse `[dis=rep-]de[-dis=rep]` de 40% `[r]` des dépenses `[dis=hes-]euh[-dis=hes]` qui restent stables `[dis=hes-]euh[-dis=hes]` en volume `[dis=hes-]euh[-dis=hes]` s'agissant des dépenses de l'état `[r]`
- `[dis=hes-]euh[-dis=hes]` contrairement à ce que vous dites le pouvoir d'achat des fonctionnaires est indexé sur l'inflation donc il n'est pas écorné il y aura des suppressions d'emplois `[r]` dans la fonction publique en vertu de la règle du un pour deux mais il n'y a pas de baisse du pouvoir euh d'achat des fonctionnaires `[dis=hes-]euh[-dis=hes]` en poste
- `(/[dis=rep-]et et et[-dis=rep]` et il faut et il faut honorer la dette `(/gratien: ça serait un peu plus de 30 mille postes non renouvelés [dis=dm-]shein[-dis=dm]/c)`
- parce que `[r]` `[dis=hes-]euh[-dis=hes]` pour des raisons diverses et pas seulement liées à la crise de 2008 puisque ça remonte à loin depuis 1973 `[pron=19 100 73]` tous nos budgets ont été en déficit
- `(/sous des gouvernements de sensibilité de/)` `(/sapin: sur la dette que nous remboursons aujourd'hui n'est pas celle de 1973 [pron=19 100 73]./c)`
- certes, mais depuis 1973 `[pron=19 100 73]` nous avons une dette qui s'est nourrie perpétuellement si je puis dire `[r]` ce qui fait qu'aujourd'hui `[r]` `[dis=hes-]euh[-dis=hes]` le service de la dette
- donc les intérêts que l'état doit payer chaque année au marché financier `[dis=hes-]euh[-dis=hes]` est supérieure `[r]` aux dépenses d'éducation et ça ça peut pas continuer comme ça

Annotation en EN

Annotation en EN

- XEmacs
- Texte issu des transcriptions
- Conventions Quaero v1.25
- Types
 - pers, func, org, loc, prod, amount, time, event
- Sous-types
 - <loc.adm.town>, <time.date.abs>, <func.coll>...

Annotations EN: exemples

il y en a eu `<time.date.abs>` `<time-modifier>` **en** `</time-modifier>` `<year>`
2003 `</year>` `</time.date.abs>` ,
et puis c' est passé puis finalement même
la
les `<pers.coll>` `<name>` **socialistes** `</name>` `</pers.coll>` ont fini par dire : "
on revient pas dessus "
parce que ça fait plaisir à personne ,
parce que c' est désagréable de travailler `<amount>` `<val>` **2** `</val>` `<unit>` **ans**
`</unit>` `<qualifier>` **de plus** `</qualifier>` `</amount>` ,
mais sauf qu' il y a pas d' autre *solution ! si il y a
d()
une autre solution , c' est la suivante qu' a proposé `<pers.ind>` `<name.first>`
Martine `</name.first>` `<name.last>` **Aubry** `</name.last>` `</pers.ind>`

Contrôle qualité

Contrôle qualité

- Transcriptions
 - Contrôle sur échantillon (~15%) : 3x3 min
 - Vérification: balises, noms, segmentation, parole superposée
 - Extraction lexicque (orthographe)
- Annotations
 - Vérifications: par type d'EN, balises chevauchées, sous-composants, EN non annotées

Conclusion

- Livraisons de données:
 - Train + Dev : juin 2011
 - Test: décembre 2011
- Corpus parole (superposée)
 - Segmentation
 - Transcription
 - Extraction d'information

Questions ?