Degré d'oralité des tweets

Fanny Ducel & Fatou Sow

Dictionnaires & néologismes

Choix du sujet

- Aspect hybride des tweets (numériqué)
- Dans l'actualité, impact du numérique sur la langue
- Nuancer la dichotomie oral/écrit
- **Hypothèse** : les tweets miment l'oral sur un support écrit

Présentation des corpus

- Corpus 1, transcriptions de l'oral : ESLO2, catégorie "cinéma"
- Corpus 2, articles extraits d'Europresse, film|cinéma
- Corpus 3, tweets contenant "le film" (Tweepy)
- Corpus 4, discours politiques (Campagne2022)

	Nb d'instances	Nb de tokens	Nb de types
Transcriptions	45	39 219	3 004
Articles	19	40 381	8 381
Tweets	1 700	40 249	7 004
Discours	6	41 507	5 823

Approche par règles

- Travail de constitution de corpus équilibrés
- String-matching sur chaque corpus : marques d'oralité et non-oralité
- Expressions régulières : négations
- Étiquetage morphosyntaxique des corpus

Approche par règles : résultats (1 / 2)

- Marques d'oralité :
  - 5080 dans les transcriptions *(euh, je, ça, ouais, bah)*
  - 571 dans les articles (*je, y, ça, alors, là)*
  - 669 dans les tweets *(je, y, ça, tu, là)*
  - 835 dans les discours politiques *(je, y, ça, là)*
- Marques de non-oralité :
  - 576 dans les transcriptions *(il, se, ne, son, lui)*
  - 1535 dans les articles *(il, son, ne, se, sa)*
  - 571 dans les tweets *(il, son, se, ne, ses)*
  - 824 dans les discours politiques *(il, ne, se, son, ses)*

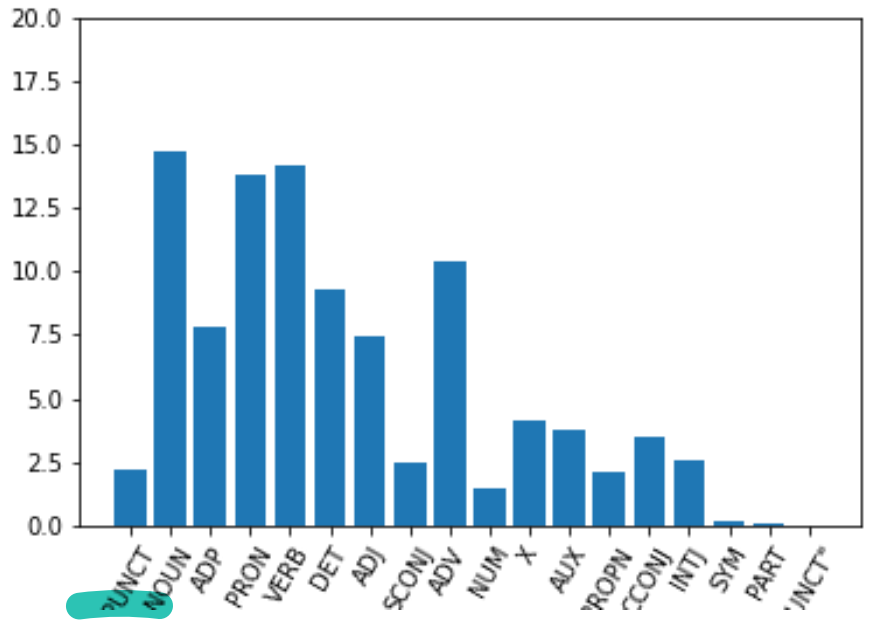
Approche par règles : résultats (1 / 2)

	Transcriptions	Articles	Tweets	Discours
% marques d'oralité	12,952	1,414	1,614	2.011
% marques de non-oralité	1,469	3,801	1,238	1.985
Négations "écrites"	51	137	32	98
Négations "orales"	280	19	89	11

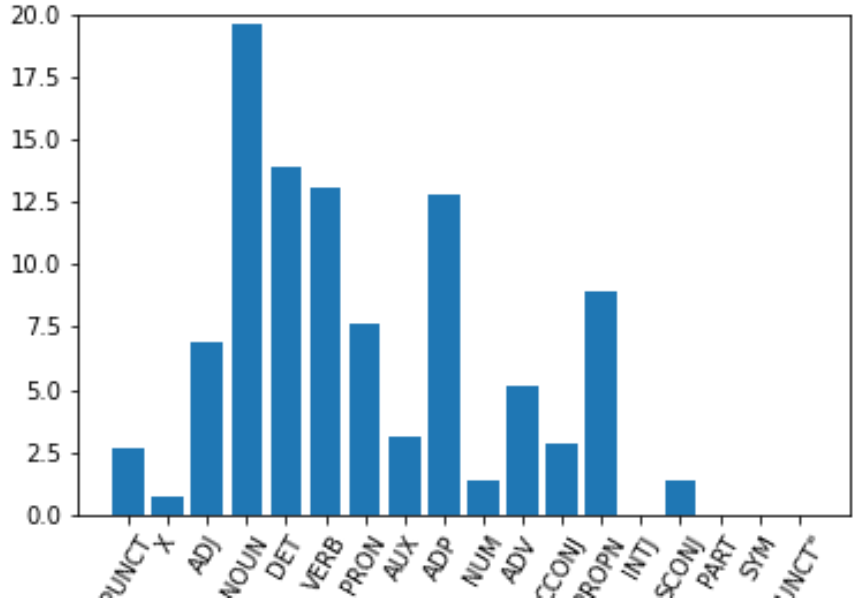
Approche par règles : POS-tagging (1 / 2)

- Stanford Pos Tagger

Transcriptions

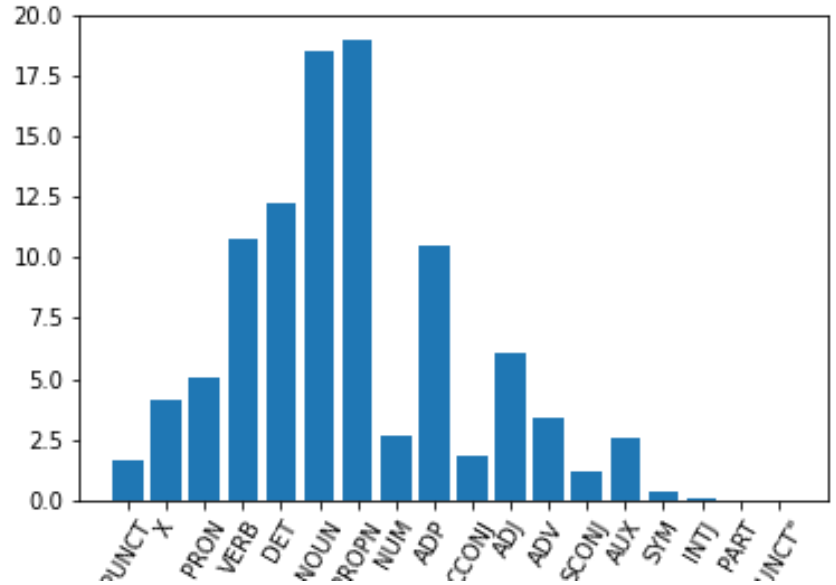


Articles

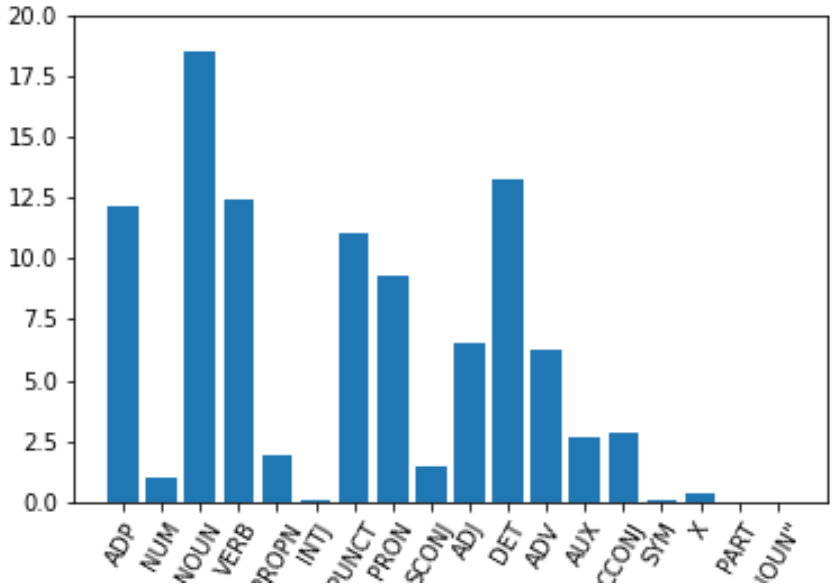


Approche par règles : POS-tagging (2/2)

Tweets



Discours politiques



Score d'oralité

- Calcul du pourcentage de chaque marque + valeur négative pour les marques de l'écrit
- Somme des valeurs des marques :
  - Si > 0 : oral
  - Si < 0 : écrit

Transcriptions	1352.51
Articles	-966.57
Discours politiques	-901.35
Tweets	-2.92

Approche automatique : classifieur non supervisé

- CountVectorizer, SVM, 0.97 d'accuracy
- Entraîné et testé sur les transcriptions et les articles étiquetés

Exemples de phrases classées orales alors qu'elles sont en réalité écrites :

- faire, même quand on se dit «*oh là là* mais *c'est pas* possible !», comme dans The Wedding Singer...
- ...Et là on s'est dit «*ah, d'accord,* ce genre de truc est possible »...
- ...Mais *on s'en fiche* ! On ne fait pas *ça* pour les gros plans, on fait *ça* pour les comédiens...
-

Exemples de phrases classées écrites alors qu'elles sont en réalité orales :

- d'aider cette femme par rapport à ce qu'elle avait vécu quoi
- ai dit effectivement c'est varié effectivement j'ai tout fait comme la miz- et après c'est pas fini comme la misère c'est presque
- hm je trouve que c'est ça qu'il manque aujourd'hui la chaleur humaine y a tellement de gens seuls y a tellement de

Classifieur sur les tweets

- Utilisé pour prédire le style (oral ou écrit) des tweets
- Résultats surprenants : **1650 étiquetés écrit, 50 oral**

Exemples de tweets classés comme oraux :

@noexpectationz Mdrrr mais si ça se trouve tu vas t'endormir devant le film tellement tu seras fatigué, c'est mieux si tu vas demain et tant pis si tu payeras +
@Louis_zoob @Sadhill71674472 @mikagx_33 @Jujutsu_anime @Crunchyroll_fr @CGR_Events Bah enfaite le film tu va voir il aide à comprendre pour la suite. Donc soit patient, et comparer one piece avec leurs derniers film sortie par rapport à jjk qui est encore a ses débuts c'est culotté, je te conseille juste d'attendre et tu verra

Exemples de tweets classés comme écrits :

RT @MarvelStory_: Vous l'attendiez, le film événement 'AVENGERS: ENDGAME' arrivera le 8 avril sur Disney+ en France !
Le film est beau, TRÈS beau, on est dans un Gotham plus sale et vivant que jamais, un Gotham plongé dans l'obscurité et la terreur. Michael Giacchino et Greig Fraser nous offrent une ambiance incroyablement oppressante et réussie. (2/5)
Le film 300 c'est trop une dinguerie
qui a vu le film ambulance ? la bande annonce est vrm stylée

Conclusion

- Articles très marqués par un style écrit, transcriptions par une forte oralité
- Discours politiques : + écrit (absence presque totale de négations orales)
- Nature des tweets : ambigu
  - Approche par règles (sauf étiquetage morphosyntaxique) : Oral
  - Approche automatique : Écrit (biais ?)

➢ Formes textuelles hybrides, autant orales qu'écrites

- Perspectives : score négation, synonymie, lexique de registre