# Reading session

Marine Wauquier

LLF Morphology Workgroup

June 11th, 2021

## Article 1

Hofmann, V., Pierrehumbert, J. & Schütze, H. (2020a). Predicting the growth of morphological families from social and linguistic factors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7273-7283.

## Article 2

Hofmann, V., Pierrehumbert, J. & Schütze, H. (2020b): DagoBERT: Generating Derivational Morphology with a Pretrained Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 16–20

- Approche computationnelle des familles morphologiques
  - Prédiction de la taille des familles
    - Évolution du nombre de dérivés pour une base donnée
    - Corpus diachronique
  - Prédiction de dérivés
    - Choix du bon procédé à partir d'une base donnée
    - Modèles de langues pré-entraînés (BERT)

- Étude des dynamiques thématiques dans les réseaux sociaux par le biais des changements lexicaux
  - Taille des familles morphologiques comme métrique complémentaire

| | |
|---|---|
| 04/2015 | trump |
| 05/2015 | trump, **trumpish**, **trumpster**, **trumpy** |
| 06/2015 | trump, **trumpening**, **trumper**, trumpish, **trumpness**, **trumpology**, trumpster, trumpy |
| 07/2015 | trump, trumpening, trumper, **trumpic**, **trumpification**, **trumpiness**, trumpish, **trupism**, **trumpistan**, trumpness, trumpster, trumpy" |

# Morphological Family Expansion Prediction

- Pour une famille morphologique donnée
  - Croissance absolue, i.e. le nombre de mots nouveaux
  - Croissance relative, i.e. le ratio de croissance
- Tâche de classification binaire
  - Binarité relative à un seuil de croissance
  - Critères linguistiques et sociaux

# Corpus

- Reddit
  - Plateforme sociale en ligne organisée en subreddits par thématique
- Échantillon de Reddit
  - 2007 à 2018
  - 4 subreddits (gaming, movies, nba, politics)
- Pré-traitement
  - Filtrage des posts issus de bots et de spammers identifiés
  - Retrait abbréviations, des formes numériques, des adresses à d'autres utilisateurs et subreddits, et liens
  - Normalisation en anglais américain
  - Réduction des répétitions de plus de 3 lettres (*niiiiice*)
  - Lemmatisation

## Famille morphologique

Soit une famille *F(w*)* dont *w** est la base (dit 'parent') et $\tilde{F}$ les dérivés morphologiques de *w** (dits 'enfants')

- Sélection des parents sur la base des 1000 mots les plus fréquents
- Segmentation des autres mots du vocabulaire
  - Liste pré-définie d'affixes
  - Prise en compte de l'allomorphie (*trumpiness* $<$ *trumpy* et non *trumpi*)
- Identification des parents candidats par algorithme récursif ascendant

# Mesure de la croissance

- Perspective diachronique
  - Sur la base des propriétés antérieures de la famille morphologique
- Découpage du corpus
  - *Probe interval*, i.e. les 6 mois à l'étude
  - *Context interval*, i.e. les 18 mois précédents
  - 68 000 paires d'intervalles étudiées

# Critères

- Nombre de formes
  - Taille moyenne de la famille sur les 18 mois précédents
- Fréquence des formes
  - Fréquence relative du parent
  - 'Trending behavior', i.e. évolution de la fréquence relative du parent entre le *context interval* et le *probe interval*
- Diffusion des formes
  - Auprès des utilisateurs
  - Au sein des fils de discussion ('thread')
  - Au sein des subreddits

- Random Forest
  - Un par croissance (absolue ou relative)
  - Un example négatif sélectionné aléatoirement pour chaque exemple positif
  - Fusion des 4 subreddits
  - Partition entre échantillons de développement et de test (80/20)
  - Seuil de croissance de 2.4 pour la croissance absolue et 1.6 pour la croissance relative

# Résultats

- Précision de 80.9% pour la croissance absolue et de 70.8% pour la croissance relative
  - Taille moyenne de la famille comme prédicteur meilleur prédicteur
  - Croissance absolue plus élevée pour les familles de grande taille
  - Croissance relative plus élevée pour les familles de petite taille
- Amélioration de la précision avec l'augmentation du seuil de croissance
  - Les changements les plus extrêmes (i.e. les croissances les plus marquées) sont plus facilement prédictibles
- 'Trending behavior' comme second meilleur prédicteur
  - Particulièrement pour les familles de petite taille

# Analyse des erreurs

- Segmentation
  - Sur 500 familles, 8.8% de mots mal appareillés
  - 60% des erreurs par 10 familles
    - represenatives (senate), heyy (hey)