Plongements de mots translingues et alignement de textes parallèles pour les dialectes alsaciens

TER Soutenance
Encadrante : Delphine BERNHARD

FALKNER Florian - 25/05/2021

Sommaire

- Contexte
- Situation linguistique
- Défis
- Enjeux
- Contributions personnelles
- Travail réalisé
- Conclusion

Contexte : les dialectes alsaciens

- Grandes variations graphiques
- Plus parlé qu'écrit :
  - Beaucoup de variété à l'écrit
  - Pas de standardisation orthographique
- Exemple de mot : aider
- Peu de ressources
- Contraintes pour le TAL
- Piste de recherche : Plongements de mots

halfa	*halfe*
helfe	*hëlfe*

***Figure 1** : Différentes formes pour le mot aider*

Situation linguistique

- *Natural Language Processing* (NLP)
  - Méthodes pour le traitement des langues
- Plusieurs dialectes alsaciens
  - Alémaniques et franciques
- Non codifiés
- Antérieurs à l'allemand standard

***Figure 2** : Carte linguistique*
https://en.wikipedia.org/wiki/ Alsatian_dialect#/media/ File:Linguistic_Map_of_Alsace.svg

Défis

- Pas de norme à l’écrit
- Impact :
  - Variation dialectale
  - Absence de norme orthographique
- Différent du français
- Petit corpus peut contenir beaucoup de variations

Enjeux

- Projet exploratoire
- Tester, évaluer et comparer les méthodes (monolingues)
  - Peu de données
  - Variation graphique (robustesse)
- Avoir de bons résultats
- Plongements de mots translingues
- Enrichir des lexiques bilingues français-alsacien

Contributions personnelles

- Test des plongements de mots pour une langue très peu dotée qui est l’alsacien
- Extraction du texte alsacien de pièces de théâtre
- Traitement des mots coupés
- Tokénisation (découpage en tokens) de 3 corpus
- Utilisation de méthodes et outils
- Évaluation et analyse de la qualité des vecteurs
- Scripts d’évaluation (Benchmark)
- Étude pourra servir de base à une analyse translingue

Travail réalisé

Données et prétraitement

Travail réalisé

- État de l’art
- Données et prétraitement
  - Corpus **THÉATRE** : 8 pièces de théâtre
  - Corpus **OSCAR** : alémanique
  - Corpus **DIVERS** : alsacien contemporain
- Extraction du texte dans THÉATRE (format XML-TEI)
- Traitement de mots coupés

Travail réalisé

- Tokénisation des corpus

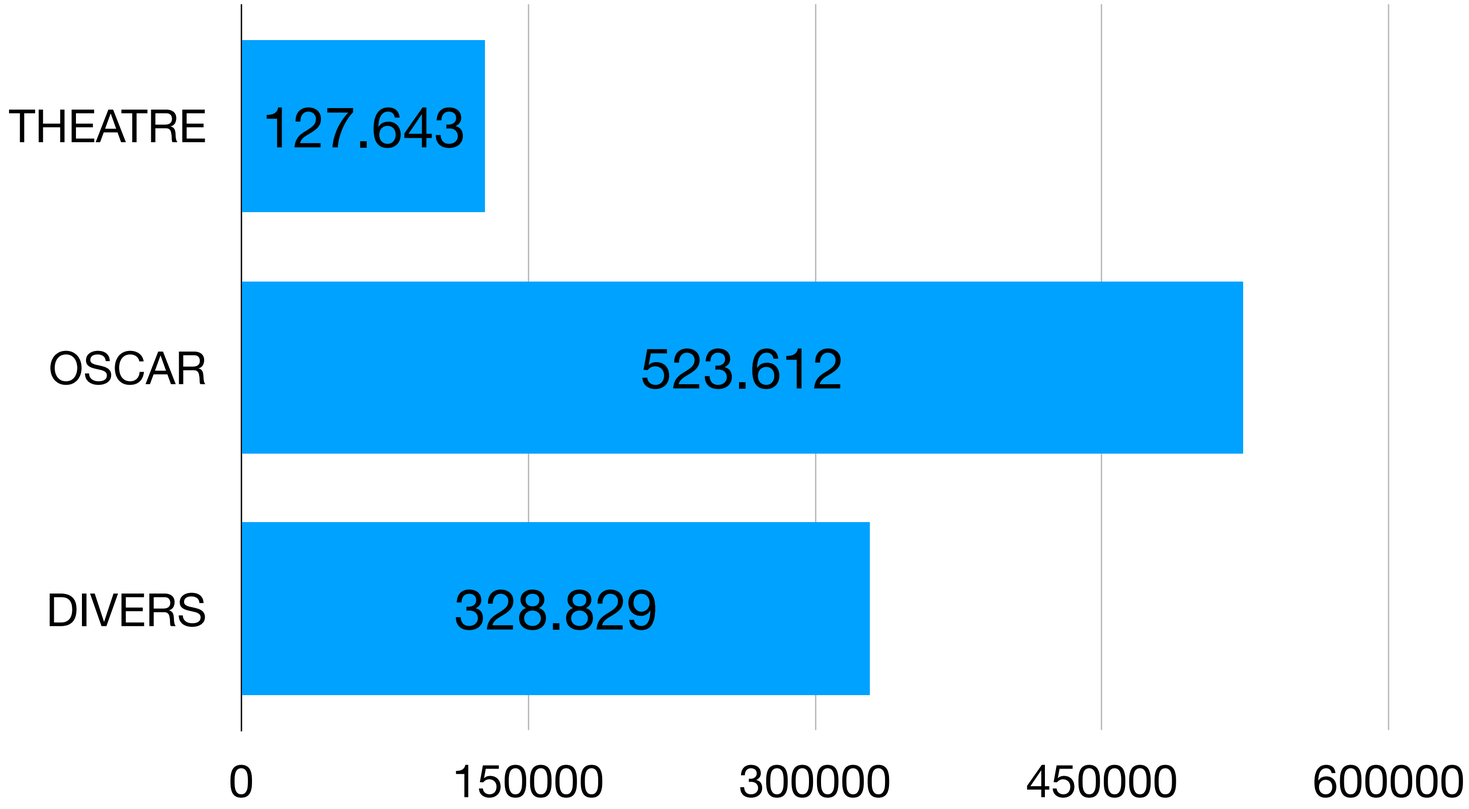

***Figure 3** : Nombre de tokens par corpus*

Travail réalisé

Plongements de mots

Travail réalisé

- Utilisation de plongements de mots
- Mot en vecteurs de réels
- Exemple : chat [0.23432, -3.23421, 1.23644, …]
- Mot graphiquement similaire à un mot ou contexte similaire

  ↳ Vecteur sera assez proche
- Exemple : chien et chat
- Calcul mathématique sur les vecteurs

Travail réalisé

Méthodes et outils

Travail réalisé

- 2 grands outils pour les plongements de mots (non-supervisé)
- **FastText** : n-grammes, mots hors vocabulaire
- **Word2Vec** : sémantique et plus rapide, possibilité d'ajouter Magnitude
- CBOW (*continuous bag-of-words*)
  - Prédit le mot en fonction de son contexte
- Skipgram
  - Meilleurs résultats
  - Prédit les mots du contexte à partir d'un mot central

Travail réalisé

- Exemple : Ma voiture est **belle** et rouge

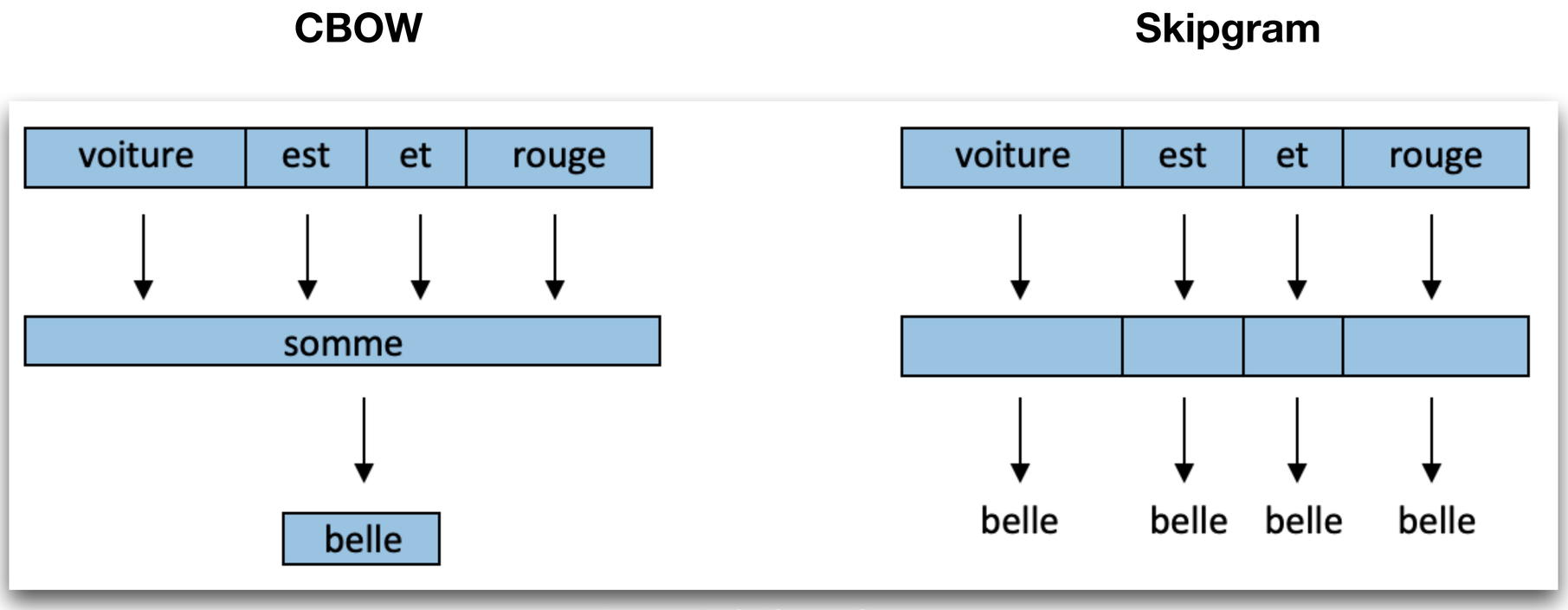

***Figure 4** : CBOW et Skipgram*

Travail réalisé

Scripts d’évaluation

Travail réalisé

- Aucun Benchmark pour l'alsacien
- Catégories de mots fréquents
- Fonctionnalités FastText :
  - **model.get_analogies**("König", "Mann", "Frau")
    - Calcul mathématique avec vecteurs : roi-homme+femme = reine
  - **model.get_nearest_neighbors**("Frau")
    - Renvoie une liste avec les 10 mots les plus proches du mot

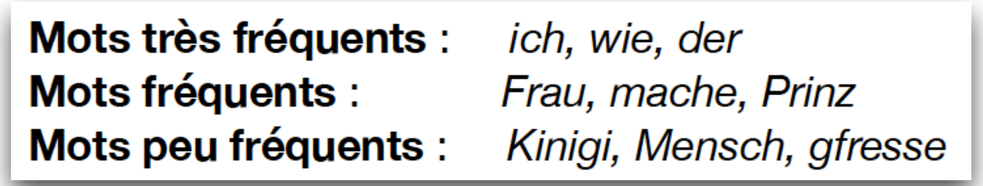

***Figure 5** : Catégories de mots*

Travail réalisé

- Lexique bilingue + Listes d’évaluation (Delphine BERNHARD)
- Catégories:
  - Synonymes morphologiquement liés - Autres synonymes
  - Variantes graphiques et flexionnelles - Même famille morphologique
  - Autre relation sémantique - Aucune relation
- Mesure de la similarité cosinus
- Mesure 10 plus proches voisins

Travail réalisé

Résultats

Travail réalisé

1) Quels modèles et architectures sont les meilleurs pour l’alsacien ?

Travail réalisé

FastText modèles

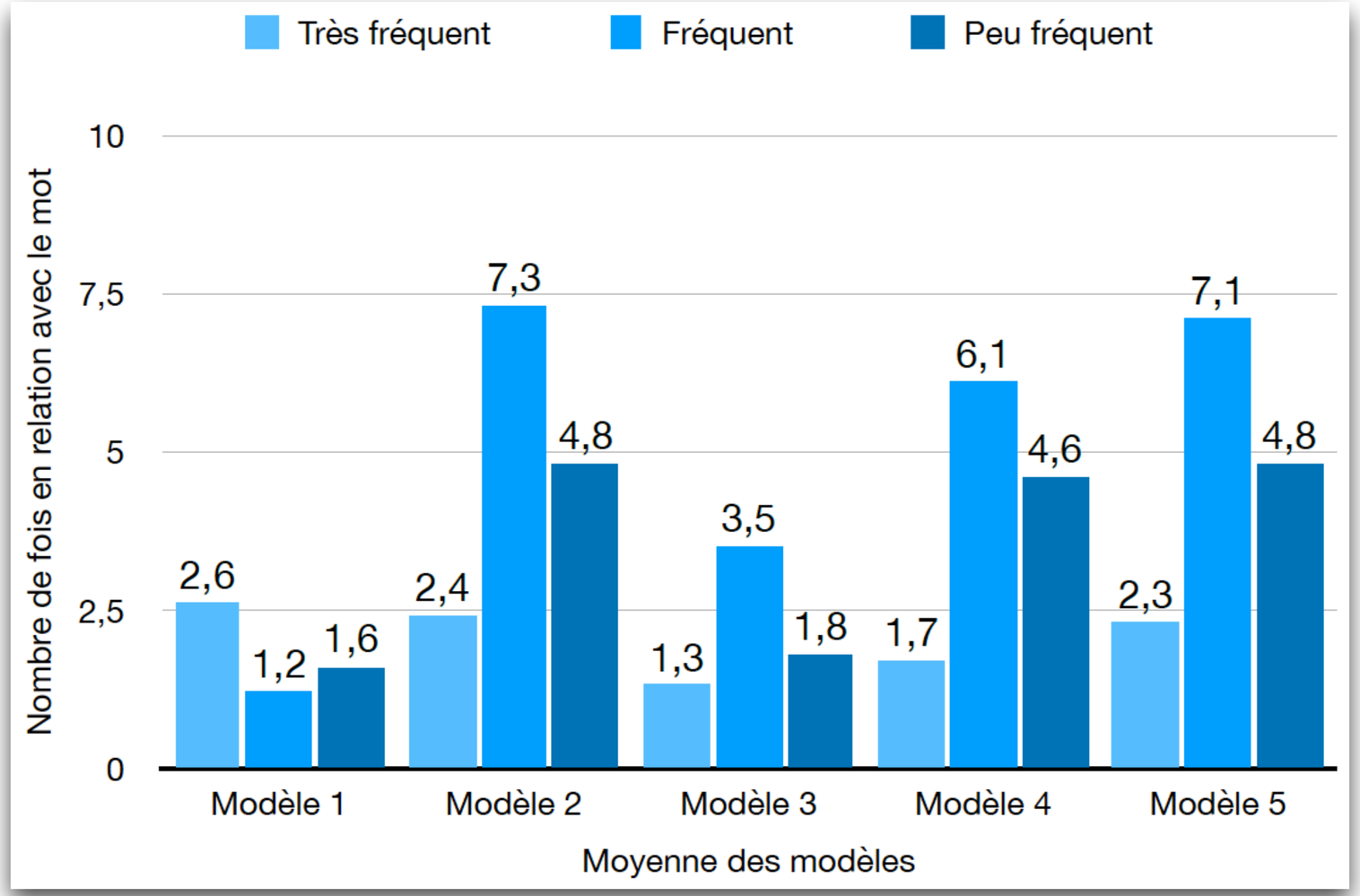

***Figure 5** : FastText comparaison*

Travail réalisé

Word2Vec avec/sans Magnitude

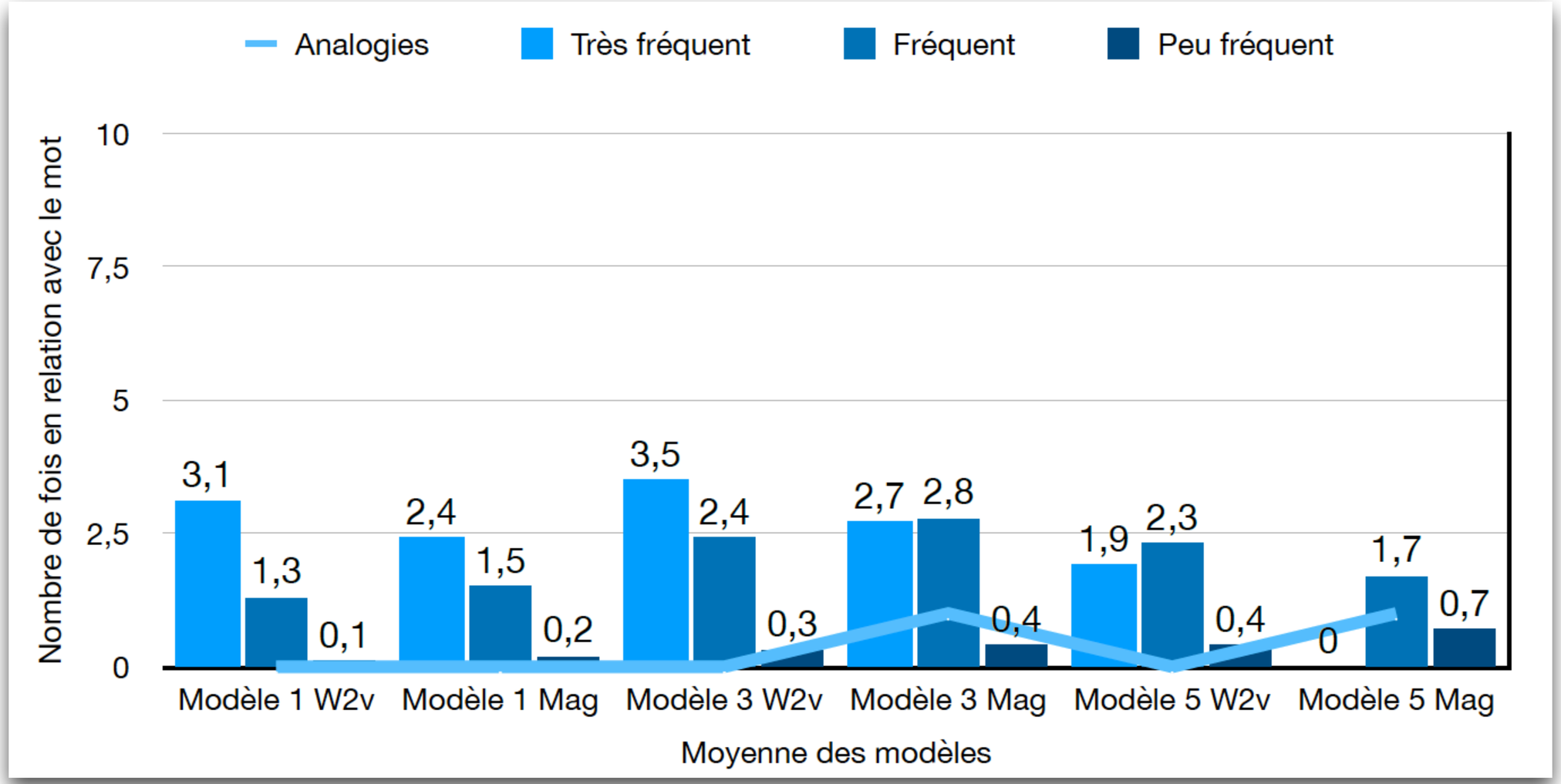

***Figure 6** : Word2Vec*

Travail réalisé

Cbow vs Skipgram

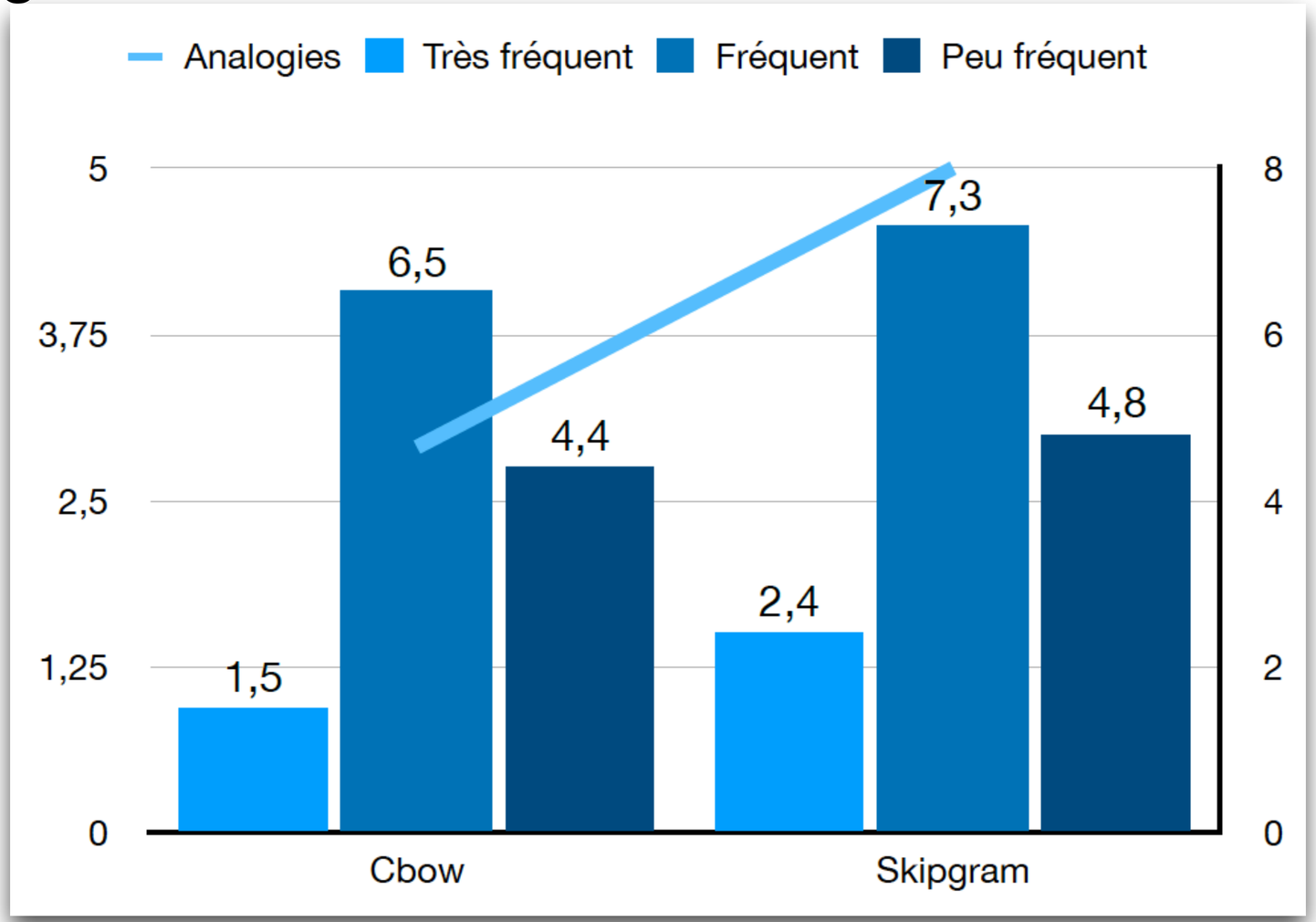

***Figure 7** : Cbow vs Skipgram*

Travail réalisé

Le modèle à retenir est :

FastText

L’architecture à retenir est :

Skipgram

Travail réalisé

2) Est-il préférable d’utiliser uniquement des textes alsaciens qui ont comme conséquence qu’il y ait moins de ressources ou vaut-il mieux faire un mélange pour avoir plus de ressources à notre disposition ?

Travail réalisé

1 corpus alémanique vs 2 corpus alsaciens vs corpus mélangés

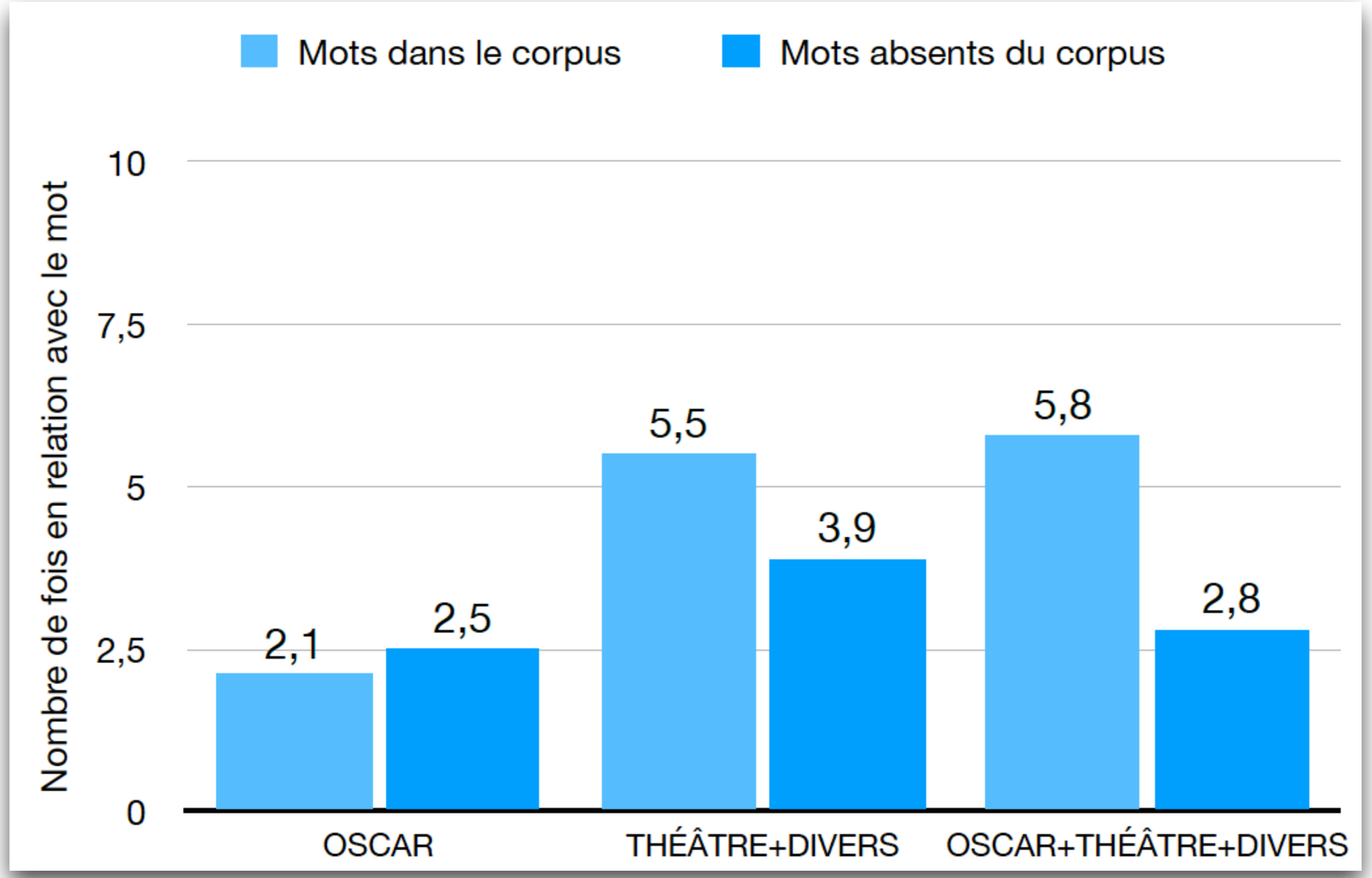

***Figure 8** : Résultats lexique bilingues sur les différents corpus*

Travail réalisé

Résultats catégories par corpus

Catégorie	Score moyen sur 300 (OSCAR+THÉÂTRE+DIVERS)	Score moyen sur 300 (THÉÂTRE+DIVERS)
Synonymes morphologiquement liés	5,4	7,4
Autres synonymes	0	1,4
Variantes graphiques et flexionnelles	45,2	54
Même famille morphologique	36	57,2
Autre relation sémantique	0	0
Mots avec aucune des relations précédentes	213,4	180,6
Score moyen - (nombre de mots pertinents)	86,6	119,2

***Figure 9** : Score moyen global de la mesure des 10 mots les plus proches*

Travail réalisé

FastText sur 2 corpus alsaciens vs corpus mélangés

```
Catégorie - Synonymes morphologiquement liés: 7/180
Catégorie - Autres synonymes: 0/180
Catégorie - Variantes graphiques et flexionnelles: 37/180
Catégorie - Même famille morphologique: 36/180
Catégorie - Autre relation sémantique: 0/180
Mots avec aucune des relations précédentes: 100/180
Score moyen - (nombre de mots pertinents): 80/180
```

Figure 10a *: Corpus alsacien (mots dans vocabulaire)*

```
Catégorie - Synonymes morphologiquement liés: 6/180
Catégorie - Autres synonymes: 0/180
Catégorie - Variantes graphiques et flexionnelles: 29/180
Catégorie - Même famille morphologique: 26/180
Catégorie - Autre relation sémantique: 0/180
Mots avec aucune des relations précédentes: 119/180
Score moyen - (nombre de mots pertinents): 61/180
```

Figure 11a *: Corpus mélangés (mots dans vocabulaire)*

```
Catégorie - Synonymes morphologiquement liés: 1/120
Catégorie - Autres synonymes: 1/120
Catégorie - Variantes graphiques et flexionnelles: 17/120
Catégorie - Même famille morphologique: 21/120
Catégorie - Autre relation sémantique: 0/120
Mots avec aucune des relations précédentes: 80/120
Score moyen - (nombre de mots pertinents): 40/120
```

Figure 10b *: Corpus alsacien (mots hors vocabulaire)*

```
Catégorie - Synonymes morphologiquement liés: 0/120
Catégorie - Autres synonymes: 0/120
Catégorie - Variantes graphiques et flexionnelles: 17/120
Catégorie - Même famille morphologique: 11/120
Catégorie - Autre relation sémantique: 0/120
Mots avec aucune des relations précédentes: 92/120
Score moyen - (nombre de mots pertinents): 28/120
```

Figure 11b *: Corpus mélangés (mots hors vocabulaire)*

Conclusion

- Cette étude :
  - Rôle important du corpus et de la méthode choisie
  - Servir de base à une analyse translingue
- Modèle stable trouvé
- Gros corpus n’est pas toujours le meilleur
- Corpus plus petit de la même famille du dialecte plutôt que des corpus mélangés

Des questions ?

Travail réalisé

FastText vs Word2Vec 2 corpus alsaciens

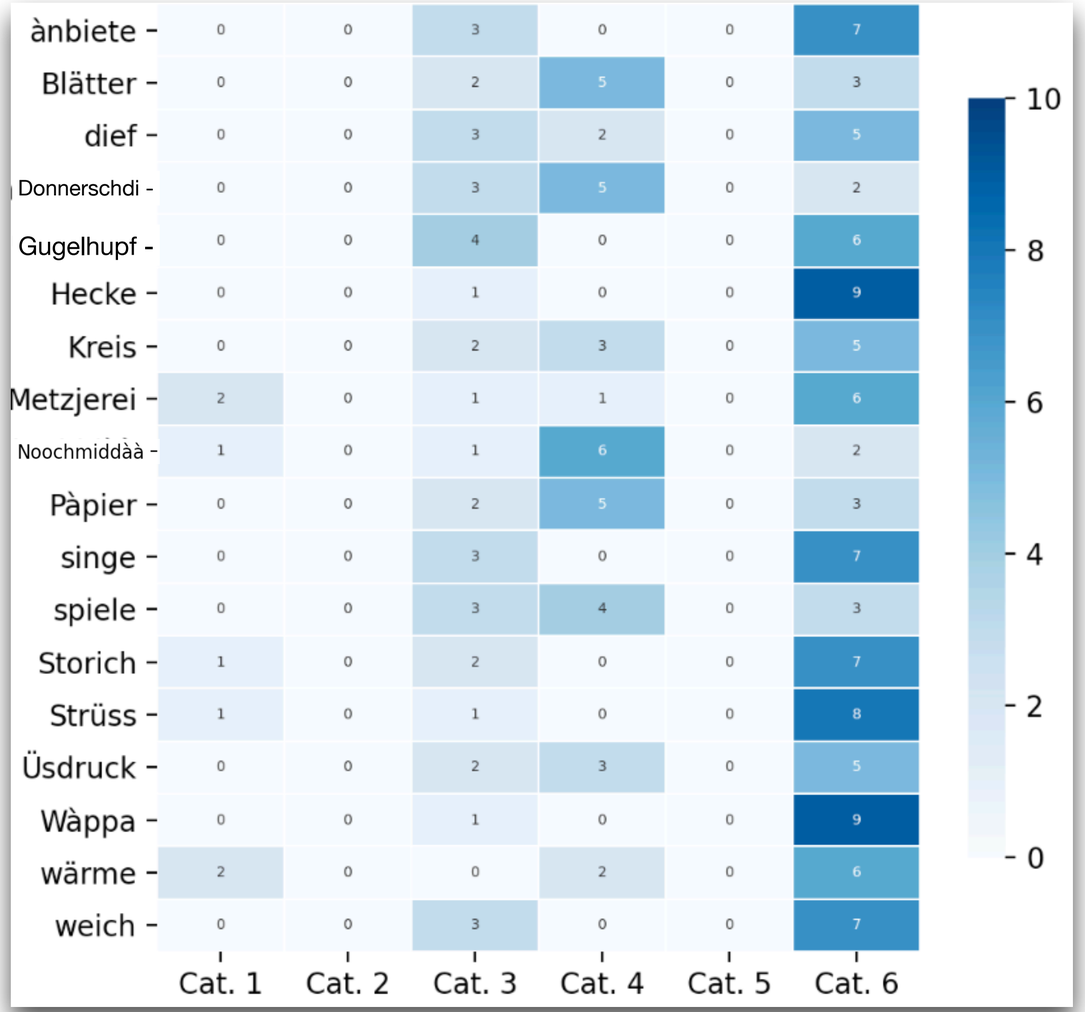

Figure 12a *: FastText*

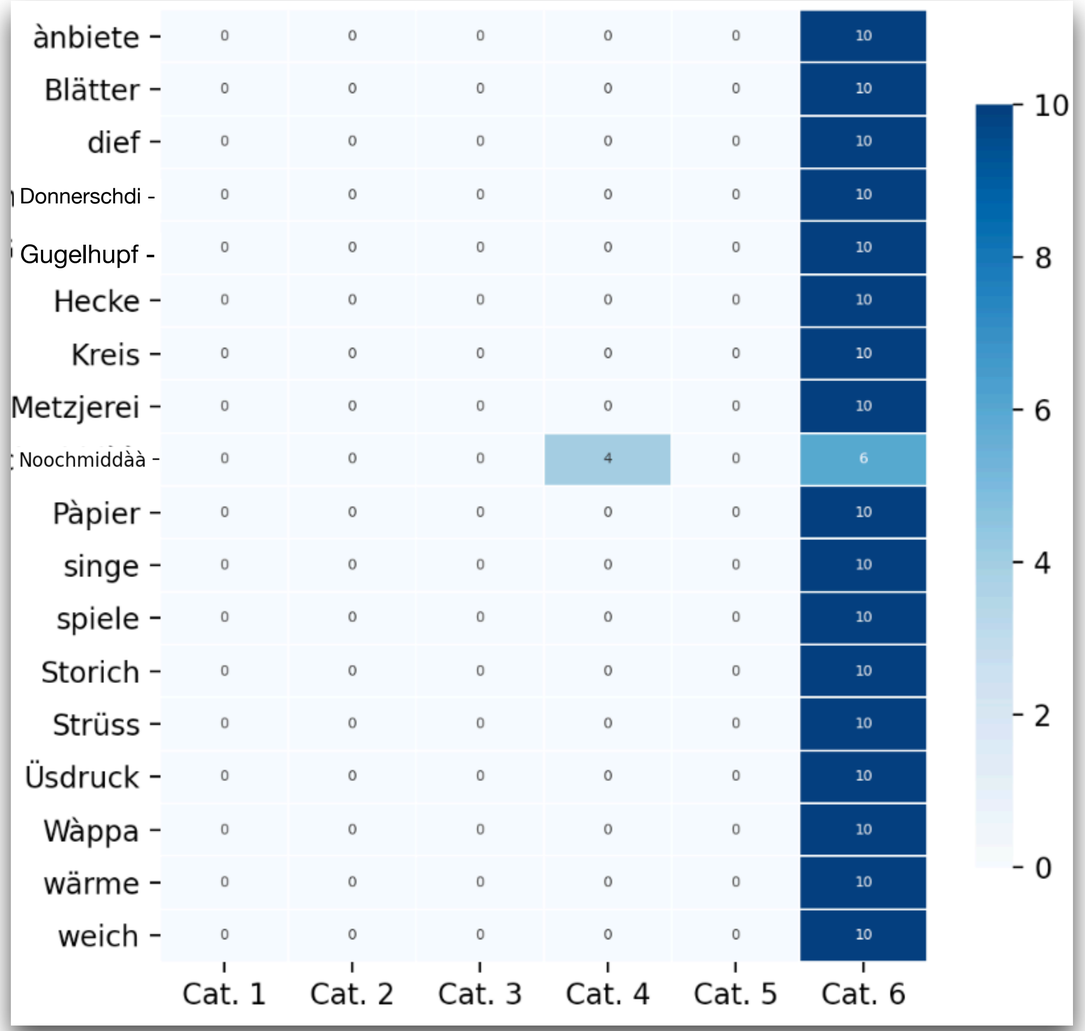

Figure 12b *: Word2Vec*