# Word Embeddings : Bénéfices d'une évaluation qualitative

Bénédicte Pierrejean

Directeur de recherche : Ludovic Tanguy

UT2J - CLLE-ERSS

20 Novembre 2017

# Sémantique distributionnelle

## Sémantique distributionnelle

Des mots apparaissant dans contextes similaires ont des sens proches (Harris, 1954)
"You shall know a word by the company it keeps", (Firth, 1957)

## Modèles distributionnels

- Modèles qui permettent de représenter les mots
- Modèles à base de fréquences
  - Construits à partir des fréquences des cooccurences
- Modèles prédictifs / dense vectors / word embeddings / plongements de mots
  - Initialisation au hasard puis ajustement pour que le vecteur ressemble à celui de ses voisins

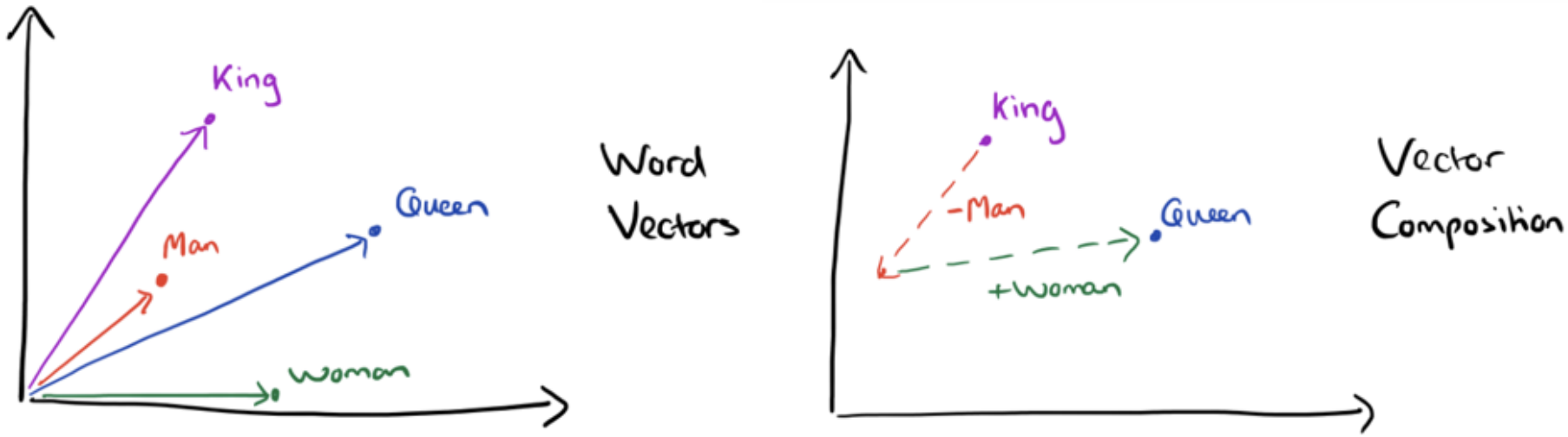
# Word embeddings : rappel


*https ://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*

- Représentation du sens des mots sous forme de vecteurs
- Pas des dimensions qu'on peut interpréter
- Modèle contenant *apple*, *microsoft* et *banana*, certaines dimensions de *apple* et *microsoft* seront similaires, d'autres pour *apple* et *banana*

# Le pouvoir des word embeddings

Différentes opérations possibles entre ces vecteurs (addition, soustraction etc.)




*https ://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*

# Analogies, autres exemples (Mikolov et al., 2013)

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

| Relationship | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| France - Paris | Italy: Rome | Japan: Tokyo | Florida: Tallahassee |
| big - bigger | small: larger | cold: colder | quick: quicker |
| Miami - Florida | Baltimore: Maryland | Dallas: Texas | Kona: Hawaii |
| Einstein - scientist | Messi: midfielder | Mozart: violinist | Picasso: painter |
| Sarkozy - France | Berlusconi: Italy | Merkel: Germany | Koizumi: Japan |
| copper - Cu | zinc: Zn | gold: Au | uranium: plutonium |
| Berlusconi - Silvio | Sarkozy: Nicolas | Putin: Medvedev | Obama: Barack |
| Microsoft - Windows | Google: Android | IBM: Linux | Apple: iPhone |
| Microsoft - Ballmer | Google: Yahoo | IBM: McNealy | Apple: Jobs |
| Japan - sushi | Germany: bratwurst | France: tapas | USA: pizza |

# Word embeddings : voisins

- Score de similarité entre deux mots : similarité cosinus
- Les mots les plus proches d'un mot : les voisins sémantiques
- Voisins de *chat* : chien, chaton, lapin, oiseau etc.

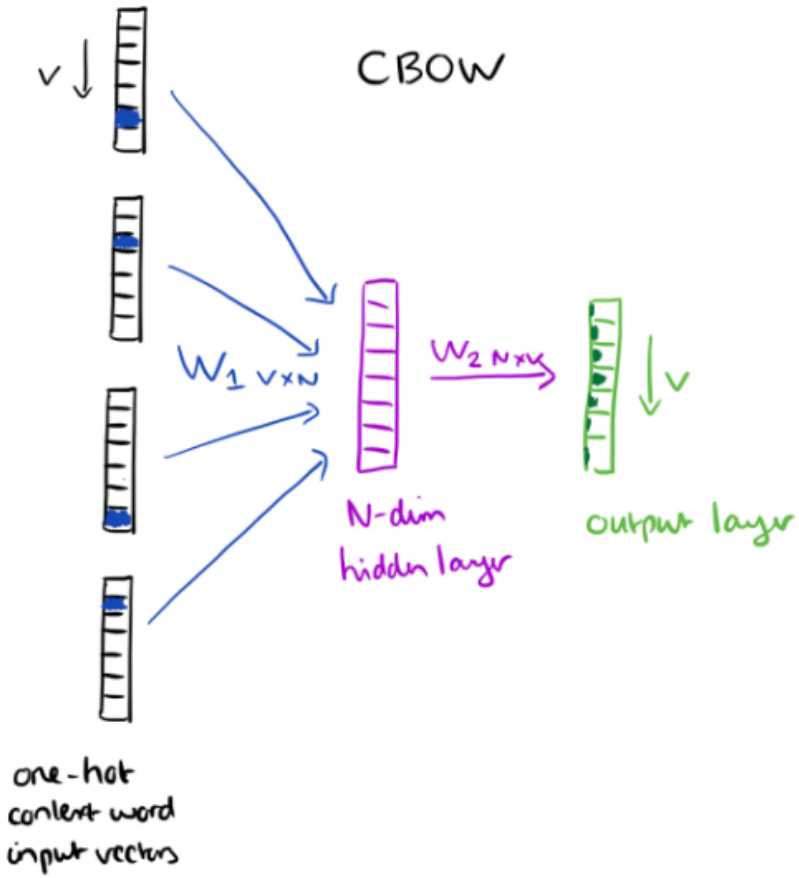
# Word2vec

## word2vec

- Mikolov et al. (2013)
- Continuous Bag of Words (CBOW)
- Skip-Gram

# CBOW

Prédiction de la probabilité qu'on mot apparaisse étant donné les mots qui l'entourent (contexte) (Colyer, 2017; Mikolov et al., 2013)



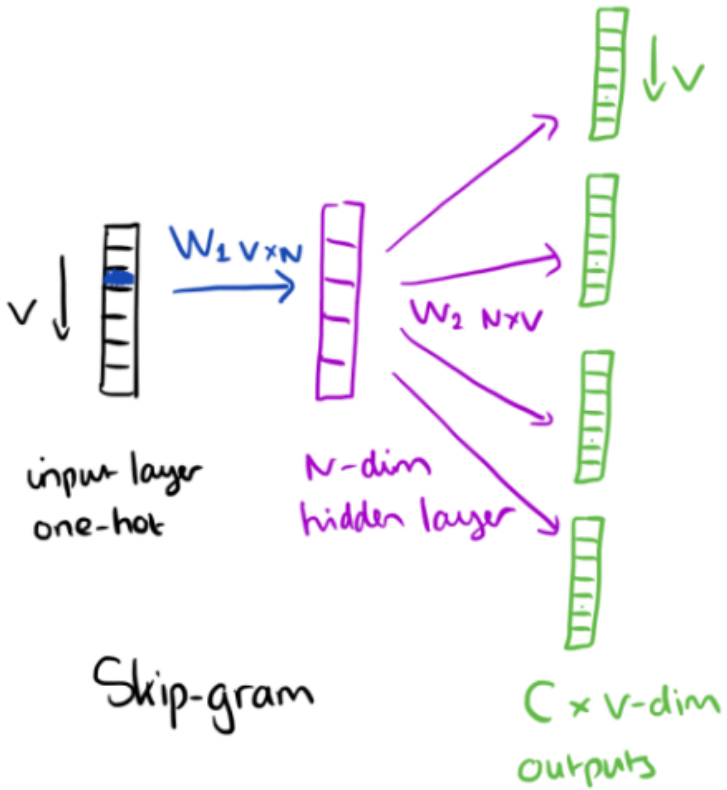

*https ://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*

# CBOW


*https ://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*

# Skip-Gram

Prédiction du contexte (Colyer, 2017; Mikolov et al., 2013)




*https ://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*

# Adaptation de word2vec

## word2vecf

- Levy and Goldberg (2014)
- Adaptation du Skip-Gram de Mikolov et al. (2013)
- Possibilité de choisir le type de contextes donnés à l'algorithme
  - Utilisation de contextes syntaxiques

# Avantages des embeddings

- Faciles à entraîner du moment qu'on a un corpus assez conséquent
- Peu de manips à faire pour avoir un modèle qui marche bien
- Rapide / efficace
- Facilement intégrable en entrée de systèmes de Deep Learning par exemple

# Quelques inconvénients

- Modèles opaques : les dimensions sont difficilement interprétables
- Plus difficile de contrôler les paramètres utilisés pour entraîner les modèles

# Plan

# Evaluation des embeddings

- Qu'est ce qui fait un "bon" embedding ?
- Représentation entre deux vecteurs doit représenter la relation linguistique entre deux mots

## Types d'évaluation

- Extrinsèque
- Intrinsèque

# Evaluation extrinsèque

Embeddings utilisés dans une tâche dont on va évaluer la performance

- étiquetage syntaxique
- reconnaissance d'entités nommées

## Avantages

- Evaluation sur des applications réelles de l'usage des embeddings
- Evaluation davantage qualitative : différents aspects vont être évalués (qualités syntaxiques, sémantiques etc.)

# Evaluation extrinsèque

## Inconvénients

- Coûteux
- Difficile à mettre en place
- Evalué pour tâche spécifique

# Evaluation intrinsèque

Test des relations sémantiques et syntaxiques entre les mots (Schnabel et al., 2015)

- Méthode la plus répandue
- Nombreux jeux d'évaluation / benchmarks
- Peu coûteux
- Bonne estimation d'un modèle qui fonctionne (ou pas)

# Exemple de benchmark : WordSim-353 (Finkelstein et al., 2002)

- 353 paires
- Degré de similarité évalué par humains
- 13 personnes
- Degré de similarité estimé sur une échelle de 0 à 10
- Possibilité d'utiliser un dictionnaire pour les mots dont le sens n'est pas connu
- Jugement individuel

# Extrait du benchmark WordSim-353

| Word1 | Word2 | Human (mean) |
|---|---|---|
| tiger | cat | 7.35 |
| book | paper | 7.46 |
| computer | keyboard | 7.62 |
| bread | butter | 6.19 |
| cucumber | potato | 5.92 |
| doctor | nurse | 7 |
| professor | doctor | 6.62 |
| king | cabbage | 0.23 |
| king | queen | 8.58 |
| professor | cucumber | 0.31 |

# Evaluation intrinsèque

## Calcul de la similarité sémantique

- Comparaison des représentations sémantiques dans notre modèle aux jugements humains.
- Similarité cosinus entre deux vecteurs
- Classement des scores des paires
- Comparaison classement ordonné de nos embeddings vs. benchmark
- Corrélation de Spearman

# Problèmes soulevés par l'évaluation intrinsèque (Baroni and Lenci, 2011; Faruqui et al., 2016)

Benchmarks pas créés spécialement pour ce type d'évaluation

## Similarité vs. Association (Faruqui et al., 2016; Baroni and Lenci, 2011)

- *Cup* vs. *coffee* considérés plus similaires que *car* et *train* dans WS-353
- Problème observé dans différents jeux d'évaluation
- WS-353 divisé en 2 jeux d'évaluation (similarité et association)
  - Qu'est ce qu'on attend du modèle ?
- Variété au sein du jeu d'évaluation
  - 1 paire de mots identiques, 17 synonymes, 28 paires hyperonymes/hyponymes, 30 paire co-hypo, 6 paires holonymes/méronymes, 92 paires association

# Problèmes soulevés par l'évaluation intrinsèque (Baroni and Lenci, 2011; Faruqui et al., 2016)

## Overfitting (Faruqui et al., 2016)

- On utilise tout le jeu d'évaluation pour entrainer, tester, rentrainer, re-tester etc.
- Jeux d'évaluation assez courts, difficile de diviser en test set/dev set.

## Éviter le surapprentissage

Ajustements sur benchmark puis test sur autre tâche, test sur plusieurs tâches pour un même modèle etc.

# Problèmes soulevés par l'évaluation intrinsèque (Baroni and Lenci, 2011; Faruqui et al., 2016)

## Polysémie

- *Bank* dans WS-353, forte similarité avec *money* (Faruqui et al., 2016)
- Pas d'infos comme la POS

Pourquoi les modèles sont différents ?

# RepEval 2017 : The Second Workshop on Evaluating Vector Space Representations for NLP

This workshop deals with the **evaluation of general-purpose vector representations** for linguistic units (...). What distinguishes these representations (or embeddings) is that they are **not trained with a specific application in mind**, but rather to **capture broadly useful features** of the represented units. (...) The **embeddings are trained with one objective, but applied on others**. Evaluating general-purpose representation learning systems is fundamentally difficult. They can be trained on a variety of objectives, making **simple intrinsic evaluations useless as a means of comparing methods**. They are also meant to be applied to a variety of downstream tasks, which will place different demands on them, making **no single extrinsic evaluation definitive**.

# BLESS (Baroni and Lenci, 2011)

- Déterminer le degré de similarité des mots proches dans l'espace sémantique
- Donner un aperçu des types de relations sémantiques générées par le modèle évalué

# BLESS (Baroni and Lenci Evaluation of Semantic Spaces)

## Concepts

- 200 concepts concrêts
- Pas de mots composés
- Mots au singulier
- Entités vivantes et non-vivantes

## Relations sémantiques

- Répartition égale des relations sémantiques
- 5 relations sémantiques : co-hyponymie (coord), hyperonymie (hyper), méronymie (mero), attribut (attri), événement (event)
- Relations *random*

# BLESS (Baroni and Lenci Evaluation of Semantic Spaces)

## Relata

- Noms, verbes, adjectifs
- Sélectionnés grâce à différentes ressources (WordNet, ConceptNet, Wikipédia etc.)
- Mots simples
- Crowdsourcing pour la validation des termes *random*

# Extrait de BLESS

| Concept | Relation | Relatum |
|---|---|---|
| restaurant | attri | clean |
| restaurant | coord | castle |
| restaurant | event | eat |
| restaurant | hyper | building |
| restaurant | mero | menu |
| restaurant | random-j | grammatical |
| restaurant | random-n | muscle |
| restaurant | random-v | play |

# Plan

# Comment évaluer des embeddings générés avec des paramètres différents ?

## Quelle influence des paramètres sur les embeddings générés ?

- Comment évaluer ce qui diffère dans ces modèles ?
- Comprendre ce qui change dans les modèles
- Observer le changement à un niveau linguistique

- Tanguy et al. (2015)
- Bernier-Colborne and Drouin (2016)

# Corpus utilisés

- 3 corpus
- Anglais
- Tailles et genres différents (spécialisé, non spécialisé)

## British National Corpus (BNC)

- 100 millions de mots
- Extraits de textes écrits et parlés fin 20ème siècle
- Sélection des textes écrits seulement
- Environ 90 millions de mots

# Corpus utilisés

## ACL Anthology Reference Corpus

- Articles scientifiques en TAL
- 22 878 articles
- Environ 100 millions de mots

## UMBC WebBase Corpus

- Plus de 3 milliards de mots
- Issu du web (Stanford WebBase project)
- Corpus nettoyé

# Génération des embeddings

## Pré-traitement des corpus

- Assemblage/mélange des textes
- Etiquetage syntaxique, lemmatisation et analyse en dépendances syntaxiques avec Talismane (Urieli, 2013)

# Exemple de sortie de Talismane

| | | | | | |
|---|---|---|---|---|---|
| 1 | In | in | IN | TMP | 8 |
| 2 | the | the | DT | NMOD | 3 |
| 3 | past | past | NN | PMOD | 1 |
| 4 | , | , | P | P | 8 |
| 5 | various | various | JJ | NMOD | 7 |
| 6 | clustering | clustering | NN | NMOD | 7 |
| 7 | approaches | approach | NNS | SBJ | 8 |
| 8 | have | have | VBP | ROOT | 0 |
| 9 | been | be | VBN | VC | 8 |
| 10 | investigated | investigate | VBN | VC | 9 |
| 11 | in | in | IN | LOC | 10 |
| 12 | document | document | NN | NMOD | 13 |
| 13 | summarization | summarization | NN | PMOD | 11 |
| 14 | . | . | P | P | 8 |

# Triplets syntaxiques

Construction de triplets syntaxiques qui sont donnés en entrée de word2vecf.

| tête | modifieur | relation syntaxique |
|---|---|---|
| NN#approach | JJ#various | NMOD |
| JJ#various | NN#approach | NMOD-1 |
| NN#approach | NN#clustering | NMOD |
| NN#clustering | NN#approach | NMOD-1 |
| VB#investigate | NN#approach | SBJ |
| NN#approach | VB#investigate | SBJ-1 |
| NN#summarization | NN#document | NMOD |
| NN#document | NN#summarization | NMOD-1 |
| VB#investigate | NN#summarization | PMOD :in |
| NN#summarization | VB#investigate | PMOD :in-1 |

## Triplets générés

ACL $\approx$ 71 000 000 triplets
BNC $\approx$ 65 000 000 triplets
UMBC $\approx$ 2 400 000 000 triplets

## Paramètres utilisés

- Corpus
- w2v SG, w2v CBOW, word2vecf
- Dimension des vecteurs : 300
- Filtrage des mots apparaissant moins de 100 fois

# Variation des voisins

Pour deux modèles, pour chaque mot du vocabulaire commun aux 2 modèles, comparaison des N premiers voisins.

## Mesure de variation des voisins

- Modèle $M_1$ et modèle $M_2$
- Vocabulaire commun
- Pour chaque mot du vocabulaire en commun
  - Récupération des N premiers voisins dans $M_1$ et $M_2$
  - Comparaison des 2 listes de voisins obtenues
  - Proportion de variation par rapport aux éléments communs aux 2 listes

$$var^n_{M_1,M_2}(w) = 1 - \frac{|neighb^n_{M_1}(w) \cap neighb^n_{M_2}(w)|}{n}$$

# Evaluation de la variation des voisins

- Comparaison des configurations 2 à 2 pour les corpus ACL et BNC (CBOW vs. SG, SG w2v vs. SG w2vf etc.)
- Variation pour 50 premiers voisins et pour 10 premiers voisins

# Exemple de variation ACL SG vs. BNC SG, 50 voisins, noms qui varient le plus

| word | pos | word_no_pos | var_score | freq_ACL | freq_BNC | log_freq_ACL | log_freq_BNC | common_voc |
|---|---|---|---|---|---|---|---|---|
| NN:hum | NN | hum | 1 | 147 | 231 | 4.990432587 | 5.442417711 | 13147 |
| NN:diversification | NN | diversification | 1 | 124 | 344 | 4.820281566 | 5.840641657 | 13147 |
| NN:tablet | NN | tablet | 1 | 224 | 286 | 5.411646052 | 5.655991811 | 13147 |
| NN:glue | NN | glue | 1 | 582 | 559 | 6.366470448 | 6.326149473 | 13147 |
| NN:hire | NN | hire | 1 | 173 | 816 | 5.153291595 | 6.704414355 | 13147 |
| NN:envelope | NN | envelope | 1 | 174 | 1154 | 5.159055299 | 7.050989447 | 13147 |
| NN:abduction | NN | abduction | 1 | 405 | 172 | 6.003887067 | 5.147494477 | 13147 |
| NN:trainer | NN | trainer | 1 | 160 | 948 | 5.075173815 | 6.854354502 | 13147 |
| NN:candy | NN | candy | 1 | 108 | 207 | 4.682131227 | 5.332718793 | 13147 |
| NN:pruning | NN | pruning | 1 | 4882 | 235 | 8.493310251 | 5.459585514 | 13147 |
| NN:mirror | NN | mirror | 1 | 157 | 2366 | 5.056245805 | 7.768956045 | 13147 |
| NN:appraisal | NN | appraisal | 1 | 287 | 936 | 5.659482216 | 6.841615476 | 13147 |
| NN:horn | NN | horn | 1 | 235 | 684 | 5.459585514 | 6.527957918 | 13147 |
| NN:subset | NN | subset | 1 | 11786 | 247 | 9.374667665 | 5.509388337 | 13147 |
| NN:crystal | NN | crystal | 1 | 145 | 1101 | 4.976733742 | 7.003974137 | 13147 |
| NN:reorganization | NN | reorganization | 1 | 124 | 364 | 4.820281566 | 5.897153868 | 13147 |
| NN:warrant | NN | warrant | 1 | 116 | 698 | 4.753590191 | 6.548219103 | 13147 |
| NN:analyser | NN | analyser | 1 | 792 | 152 | 6.674561392 | 5.023880521 | 13147 |
| NN:shuffle | NN | shuffle | 1 | 132 | 113 | 4.882801923 | 4.727387819 | 13147 |
| NN:derivation | NN | derivation | 1 | 11307 | 196 | 9.333177282 | 5.278114659 | 13147 |
| NN:surf | NN | surf | 1 | 109 | 211 | 4.691347882 | 5.351858133 | 13147 |
| NN:portal | NN | portal | 1 | 223 | 124 | 5.407171771 | 4.820281566 | 13147 |
| NN:decomposition | NN | decomposition | 1 | 3866 | 164 | 8.25997566 | 5.099866428 | 13147 |

# Exemples de voisins pour les mots avec variation élevée

*pruning*
ACL : coarse-to-fine, thresholding, smoothing, filtering, beam
BNC : planting, sowing, weeding, digging, harvesting

*diversification*
ACL : search, clustering, graph-theoretic, query, sorting
BNC : rationalisation, innovation, specialization, development, expansion

*envelope*
ACL : lip, treillis, triangle, contour, waveform
BNC : folder, wallet, bag, postcard, briefcase

# Exemple de variation ACL SG vs. BNC SG, 50 voisins, noms qui varient le moins

| word | pos | word_no_pos | var_score | freq_ACL | freq_BNC | log_freq_ACL | log_freq_BNC | common_voc |
|---|---|---|---|---|---|---|---|---|
| NN:focus | NN | focus | 0.6 | 13144 | 3527 | 9.48372066 | 8.16820293 | 13147 |
| NN:science | NN | science | 0.6 | 6215 | 7956 | 8.734721004 | 8.98168164 | 13147 |
| NN:importance | NN | importance | 0.6 | 8143 | 9224 | 9.004913941 | 9.129564062 | 13147 |
| NN:potential | NN | potential | 0.58 | 3573 | 4205 | 8.181160858 | 8.344029572 | 13147 |
| NN:lack | NN | lack | 0.58 | 6019 | 8495 | 8.702676412 | 9.047233034 | 13147 |
| NN:morning | NN | morning | 0.58 | 921 | 14909 | 6.825460036 | 9.609720336 | 13147 |
| NN:study | NN | study | 0.58 | 24798 | 17882 | 10.11851828 | 9.791549899 | 13147 |
| NN:mismatch | NN | mismatch | 0.58 | 1251 | 135 | 7.13169851 | 4.905274778 | 13147 |
| NN:summer | NN | summer | 0.58 | 689 | 9748 | 6.535241271 | 9.184817415 | 13147 |
| NN:change | NN | change | 0.58 | 8553 | 19067 | 9.054037378 | 9.855714371 | 13147 |
| NN:dozen | NN | dozen | 0.56 | 381 | 2315 | 5.942799375 | 7.747164967 | 13147 |
| NN:increase | NN | increase | 0.56 | 6017 | 9407 | 8.702344075 | 9.149209372 | 13147 |
| NN:discrepancy | NN | discrepancy | 0.56 | 611 | 359 | 6.415096959 | 5.883322388 | 13147 |
| NN:emphasis | NN | emphasis | 0.56 | 2017 | 5171 | 7.609366538 | 8.550821372 | 13147 |
| NN:variety | NN | variety | 0.54 | 10305 | 8258 | 9.240384493 | 9.018937706 | 13147 |
| NN:connection | NN | connection | 0.54 | 3258 | 3183 | 8.088868789 | 8.065579427 | 13147 |
| NN:attempt | NN | attempt | 0.54 | 4194 | 8784 | 8.341410211 | 9.080687164 | 13147 |
| NN:indication | NN | indication | 0.54 | 1951 | 2145 | 7.576097341 | 7.670894831 | 13147 |
| NN:difference | NN | difference | 0.54 | 19281 | 9677 | 9.866875434 | 9.177507215 | 13147 |
| NN:day | NN | day | 0.54 | 4372 | 49157 | 8.382975849 | 10.80277454 | 13147 |
| NN:need | NN | need | 0.54 | 8635 | 15795 | 9.063578991 | 9.667448713 | 13147 |
| NN:everything | NN | everything | 0.5 | 1264 | 13620 | 7.142036575 | 9.51929458 | 13147 |
| NN:aim | NN | aim | 0.5 | 4437 | 4896 | 8.397733751 | 8.496173824 | 13147 |
| NN:week | NN | week | 0.48 | 1762 | 24260 | 7.474204807 | 10.09658418 | 13147 |

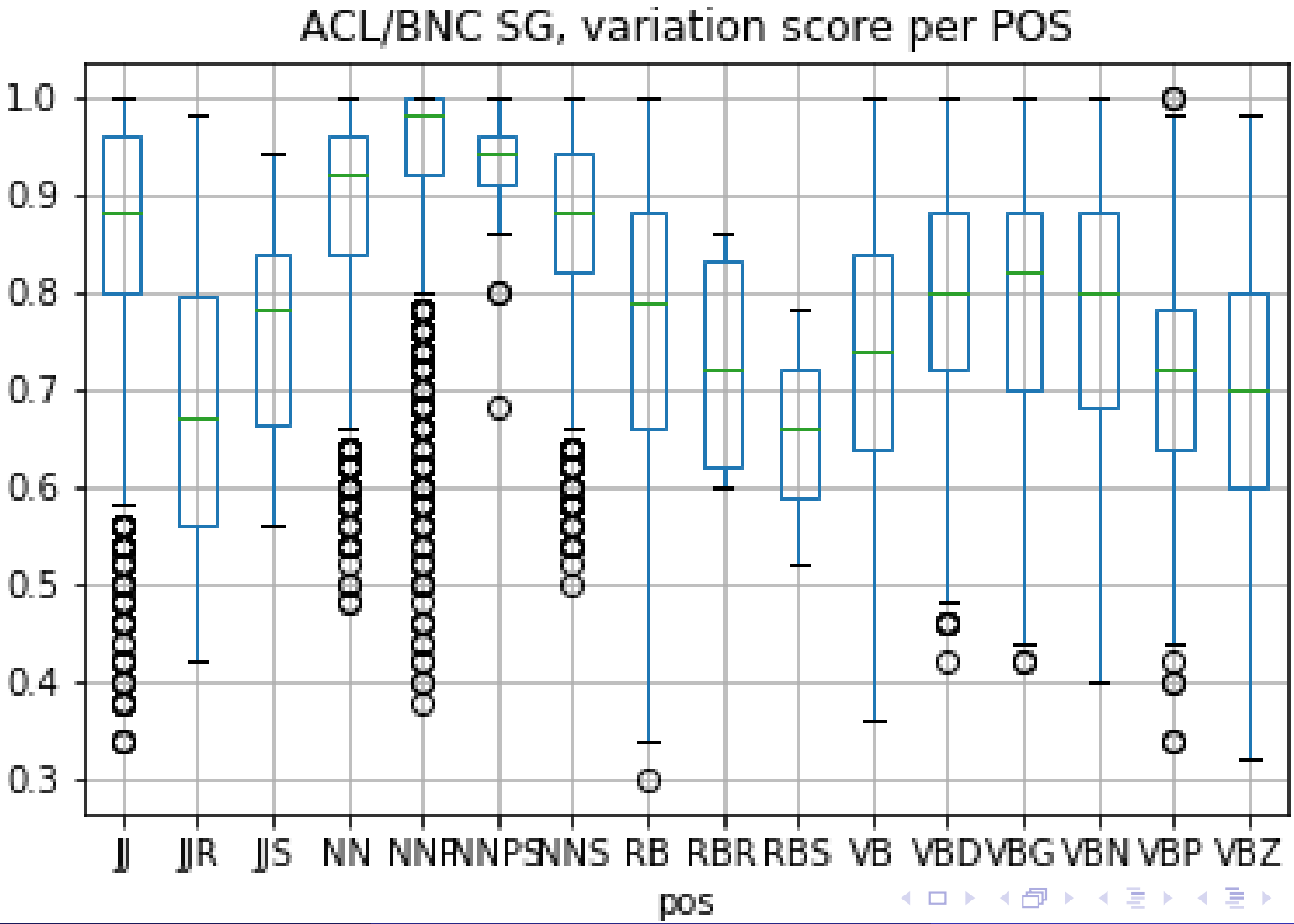
# Exemples de voisins pour les mots avec le moins de variation

*week*
month, year, season, tuesday, saturday etc.

*aim*
objective, intention, purpose, intend, focus

# Analyse des résultats

## Répartition de la variation par étiquette syntaxique



# Variation sur le corpus ACL

| | | ACL w2v SG - w2vf 100m l | ACL w2v CBOW - w2vf 100m l | ACL SG - CBOW 100m l |
|---|---|---|---|---|
| N-50 | NN var median | 0.88 | 0.84 | 0.46 |
| | NNS var median | 0.84 | 0.78 | 0.44 |
| | JJ var median | 0.84 | 0.82 | 0.44 |
| | VB var median | 0.84 | 0.8 | 0.46 |
| | VBD var median | 0.83 | 0.8 | 0.44 |
| | VBG var median | 0.88 | 0.84 | 0.44 |
| | VBN var median | 0.84 | 0.76 | 0.5 |
| | VBP var median | 0.78 | 0.74 | 0.42 |
| | VBZ var median | 0.78 | 0.74 | 0.4 |
| | mean var | 0.85 | 0.82 | 0.5 |
| | median var | 0.88 | 0.84 | 0.5 |
| | | | | |
| N-10 | NN var median | 0.9 | 0.9 | 0.5 |
| | NNS var median | 0.9 | 0.8 | 0.4 |
| | JJ var median | 0.8 | 0.8 | 0.4 |
| | VB var median | 0.9 | 0.9 | 0.5 |
| | VBD var median | 0.9 | 0.9 | 0.4 |
| | VBG var median | 0.9 | 0.9 | 0.4 |
| | VBN var median | 0.9 | 0.8 | 0.5 |
| | VBP var median | 0.8 | 0.8 | 0.4 |
| | VBZ var median | 0.8 | 0.8 | 0.4 |
| | mean var | 0.86 | 0.83 | 0.5 |
| | median var | 0.9 | 0.9 | 0.5 |
| common voc size (unique tokens) | | 23091 | 23091 | 26412 |

# Variation sur le corpus BNC

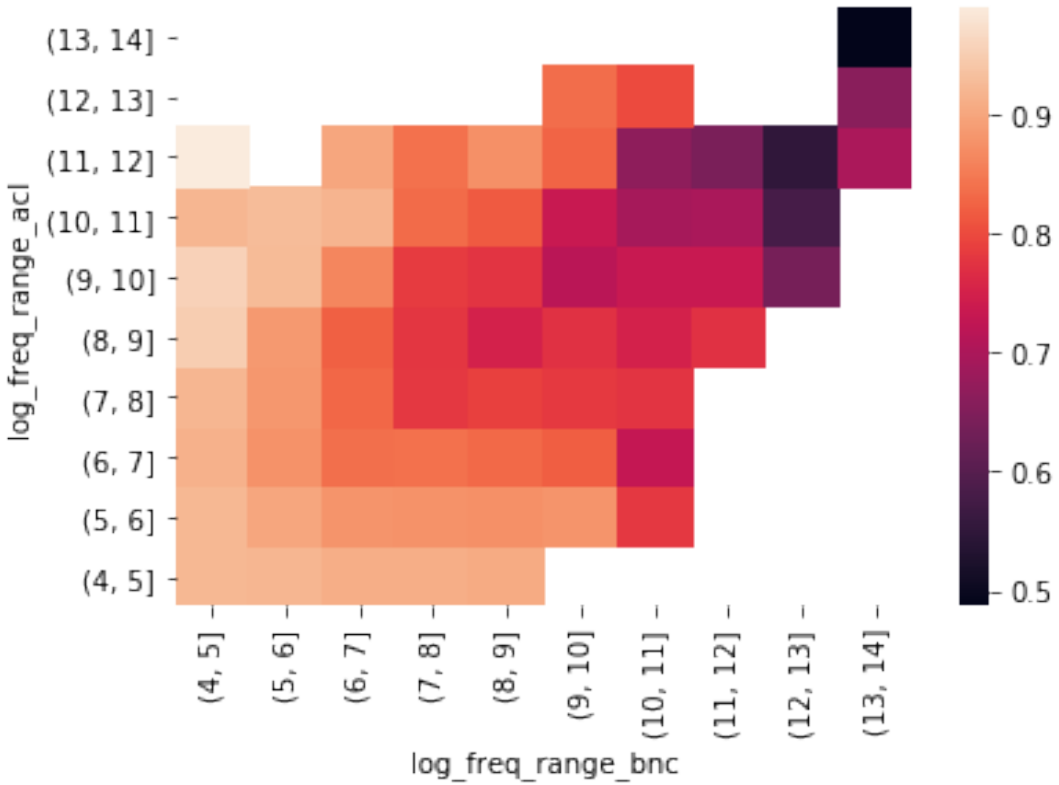
| | | BNC w2v SG - w2vf 100m l | BNC w2v CBOW - w2vf 100m l | BNC SG - CBOW 100m l |
|---|---|---|---|---|
| N-50 | NN var median | 0.9 | 0.86 | 0.44 |
| | NNS var median | 0.86 | 0.82 | 0.4 |
| | JJ var median | 0.9 | 0.86 | 0.45 |
| | VB var median | 0.88 | 0.86 | 0.5 |
| | VBD var median | 0.84 | 0.82 | 0.44 |
| | VBG var median | 0.92 | 0.9 | 0.48 |
| | VBN var median | 0.9 | 0.86 | 0.5 |
| | VBP var median | 0.8 | 0.76 | 0.42 |
| | VBZ var median | 0.84 | 0.82 | 0.42 |
| | mean var | 0.87 | 0.84 | 0.47 |
| | median var | 0.9 | 0.88 | 0.46 |
| | | | | |
| N-10 | NN var median | 0.9 | 0.9 | 0.5 |
| | NNS var median | 0.9 | 0.9 | 0.4 |
| | JJ var median | 0.9 | 0.9 | 0.5 |
| | VB var median | 1 | 0.9 | 0.6 |
| | VBD var median | 0.9 | 0.9 | 0.5 |
| | VBG var median | 1 | 1 | 0.4 |
| | VBN var median | 1 | 0.9 | 0.5 |
| | VBP var median | 0.9 | 0.9 | 0.5 |
| | VBZ var median | 0.9 | 0.9 | 0.4 |
| | mean var | 0.9 | 0.87 | 0.49 |
| | median var | 0.9 | 0.9 | 0.5 |
| common voc size (unique tokens) | | 32317 | 32317 | 34115 |

# Variation sur ACL-BNC

| | | ACL - BNC w2vf 100m l | ACL-BNC w2v SG 100m l | ACL-BNC w2v CBOW 100m l |
|---|---|---|---|---|
| N-50 | NN var median | 0.98 | 0.92 | 0.9 |
| | NNS var median | 0.96 | 0.88 | 0.84 |
| | JJ var median | 0.96 | 0.88 | 0.86 |
| | VB var median | 0.92 | 0.74 | 0.72 |
| | VBD var median | 0.96 | 0.8 | 0.8 |
| | VBG var median | 0.94 | 0.82 | 0.78 |
| | VBN var median | 0.94 | 0.8 | 0.74 |
| | VBP var median | 0.92 | 0.72 | 0.72 |
| | VBZ var median | 0.92 | 0.7 | 0.68 |
| | mean var | 0.93 | 0.85 | 0.83 |
| | median var | 0.96 | 0.88 | 0.86 |
| | | | | |
| N-10 | NN var median | 1 | 0.9 | 0.9 |
| | NNS var median | 1 | 0.8 | 0.8 |
| | JJ var median | 1 | 0.9 | 0.9 |
| | VB var median | 1 | 0.7 | 0.8 |
| | VBD var median | 1 | 0.7 | 0.8 |
| | VBG var median | 1 | 0.7 | 0.7 |
| | VBN var median | 1 | 0.7 | 0.8 |
| | VBP var median | 1 | 0.6 | 0.8 |
| | VBZ var median | 1 | 0.6 | 0.7 |
| | mean var | 0.95 | 0.8 | 0.82 |
| | median var | 1 | 0.9 | 0.9 |
| common voc size (unique tokens) | | 13989 | 13147 | 13147 |

# Influence de la fréquence sur la variation

Est-ce que la fréquence va faire qu'un mot varie plus ou moins ?



# Observation de la variation

Parmi la variation observée, qu'est-ce qui varie beaucoup, qu'est-ce qui ne varie pas ?

Parmi ce qui varie le moins, cas des adjectifs de nationalité.

# Le cas des adjectifs de nationalité : ACL

| word | ACL CBOW-SG | ACL CBOW-w2vf | ACL SG-w2vf |
|---|---|---|---|
| french | 0.22 | 0.38 | 0.4 |
| italian | 0.18 | 0.34 | 0.34 |
| swedish | 0.16 | 0.32 | 0.34 |
| chinese | 0.38 | 0.42 | 0.56 |
| dutch | 0.26 | 0.42 | 0.58 |
| spanish | 0.14 | 0.3 | 0.3 |
| turkish | 0.22 | 0.42 | 0.4 |
| belgian | 0.62 | 0.9 | 0.9 |
| russian | 0.2 | 0.46 | 0.42 |
| british | 0.52 | 0.64 | 0.74 |
| japanese | 0.3 | 0.4 | 0.56 |
| danish | 0.22 | 0.66 | 0.66 |

# Le cas des adjectifs de nationalité : BNC

| word | BNC CBOW-SG | BNC CBOW-w2vf | BNC SG-w2vf |
|---|---|---|---|
| french | 0.16 | 0.38 | 0.42 |
| italian | 0.24 | 0.38 | 0.46 |
| swedish | 0.16 | 0.44 | 0.44 |
| chinese | 0.3 | 0.52 | 0.58 |
| dutch | 0.12 | 0.46 | 0.52 |
| spanish | 0.14 | 0.5 | 0.56 |
| turkish | 0.2 | 0.38 | 0.44 |
| belgian | 0.26 | 0.5 | 0.48 |
| russian | 0.24 | 0.36 | 0.48 |
| british | 0.26 | 0.32 | 0.38 |
| japanese | 0.28 | 0.46 | 0.4 |
| danish | 0.18 | 0.46 | 0.56 |

# Le cas des adjectifs de nationalité : ACL/BNC

| word | BNC/ACL CBOW | BNC/ACL SG | BNC/ACL w2vf |
|---|---|---|---|
| french | 0.44 | 0.5 | 0.68 |
| italian | 0.52 | 0.46 | 0.64 |
| swedish | 0.56 | 0.58 | 0.74 |
| chinese | 0.72 | 0.76 | 0.62 |
| dutch | 0.46 | 0.46 | 0.62 |
| spanish | 0.46 | 0.52 | 0.72 |
| turkish | 0.68 | 0.68 | 0.7 |
| belgian | 0.64 | 0.8 | 0.92 |
| russian | 0.6 | 0.6 | 0.68 |
| british | 0.66 | 0.8 | 0.6 |
| japanese | 0.7 | 0.72 | 0.62 |
| danish | 0.54 | 0.52 | 0.74 |

# Le cas des adjectifs de nationalité : ACL/UMBC SG

| word | UMBC/ACL SG |
|---|---|
| french | 0.48 |
| italian | 0.46 |
| swedish | 0.48 |
| chinese | 0.74 |
| dutch | 0.48 |
| spanish | 0.62 |
| turkish | 0.74 |
| belgian | 0.78 |
| russian | 0.64 |
| british | 0.82 |
| japanese | 0.7 |
| danish | 0.56 |

# Voisins des adjectifs de nationalité dans ACL et BNC SG

| word | ACL | BNC | word | ACL | BNC |
|---|---|---|---|---|---|
| french | spanish | spanish | italian | spanish | spanish |
| | german | italian | | french | french |
| | italian | belgian | | portuguese | dutch |
| | english | german | | german | portuguese |
| | russian | dutch | | swedish | belgian |
| | portuguese | portuguese | | dutch | austrian |

Quels sont les contextes partagés par ces adjectifs ?

# Relations sémantiques captées par les différents modèles

Est-ce que selon le système utilisé, les paramètres, le modèle aura tendance à représenter davantage une relation sémantique plutôt qu'une autre ?

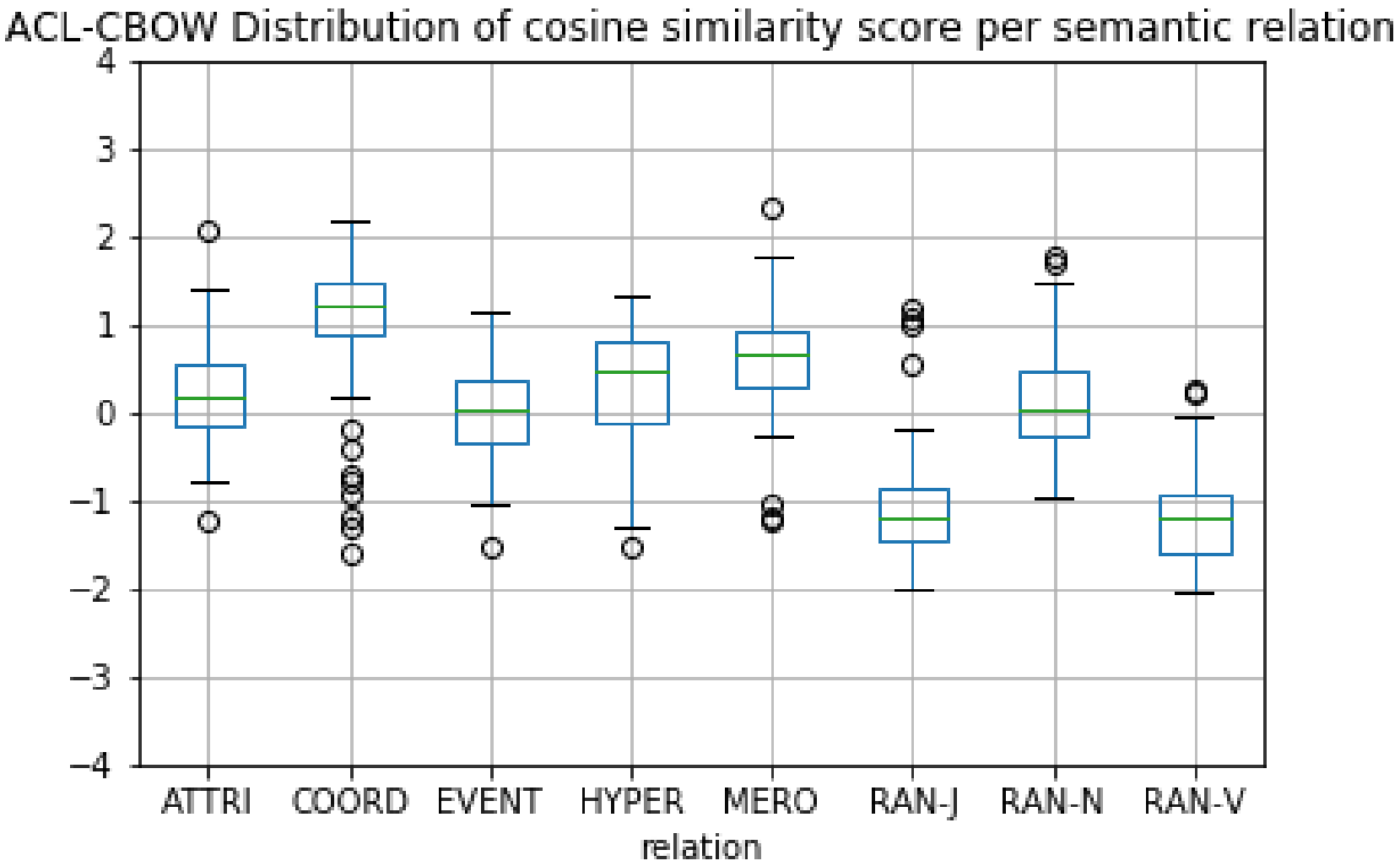
# Exemple d'évaluation avec BLESS

| target | class | relation | relatum |
|---|---|---|---|
| car-n | vehicle | attri | beautiful-j |
| car-n | vehicle | attri | big-j |
| car-n | vehicle | attri | black-j |
| car-n | vehicle | attri | cheap-j |
| car-n | vehicle | attri | clean-j |
| car-n | vehicle | attri | comfortable-j |
| car-n | vehicle | attri | dirty-j |
| car-n | vehicle | attri | dusty-j |
| car-n | vehicle | attri | expensive-j |
| car-n | vehicle | attri | fancy-j |
| car-n | vehicle | coord | ambulance-n |
| car-n | vehicle | coord | battleship-n |
| car-n | vehicle | coord | bicycle-n |
| car-n | vehicle | coord | bike-n |
| car-n | vehicle | coord | bus-n |
| car-n | vehicle | coord | ferry-n |
| car-n | vehicle | coord | fighter-n |

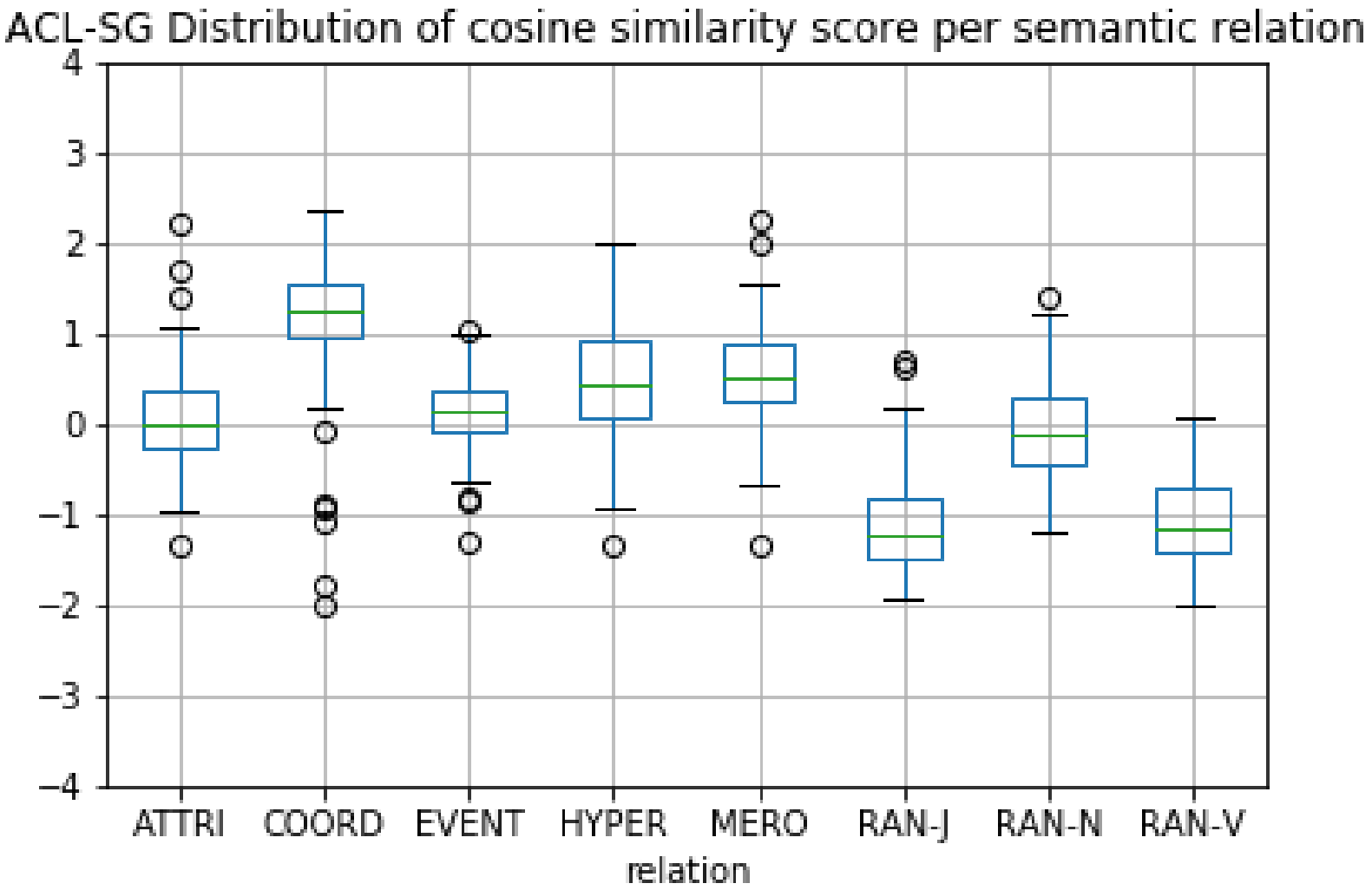
# Exemple d'évaluation avec BLESS ACL SG

| target | relation | best_relatum | best_score |
|---|---|---|---|
| NN#car | coord | NN#truck | 0.66417723 |
| NN#car | attri | JJ#dirty | 0.45889867 |
| NN#car | random-v | VB#listen | 0.26455066 |
| NN#car | hyper | NN#vehicle | 0.48279719 |
| NN#car | random-n | NN#soil | 0.45556031 |
| NN#car | event | VB#buy | 0.4236146 |
| NN#car | mero | NN#gasoline | 0.61605584 |
| NN#car | random-j | JJ#violent | 0.37985817 |
| NN#radio | coord | NN#television | 0.68221603 |
| NN#radio | attri | JJ#loud | 0.42999051 |
| NN#radio | random-v | VB#arrange | 0.23608224 |
| NN#radio | hyper | NN#equipment | 0.35459234 |
| NN#radio | random-n | NN#strike | 0.34062268 |
| NN#radio | event | VB#listen | 0.4486297 |
| NN#radio | mero | NN#music | 0.43336856 |
| NN#radio | random-j | JJ#professional | 0.26108893 |

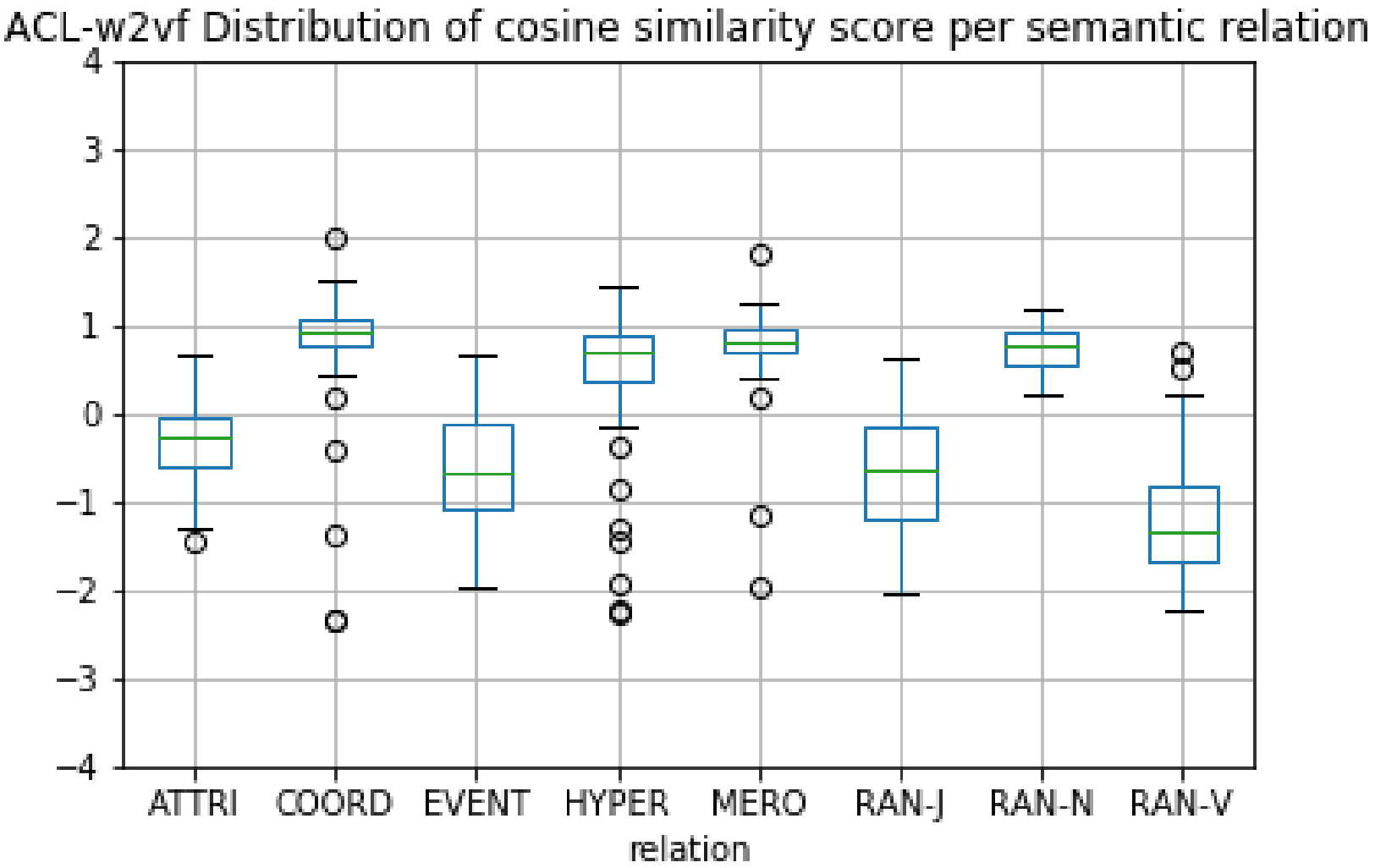
# Evaluation à l'aide de BLESS : ACL CBOW



# Evaluation à l'aide de BLESS : ACL SG



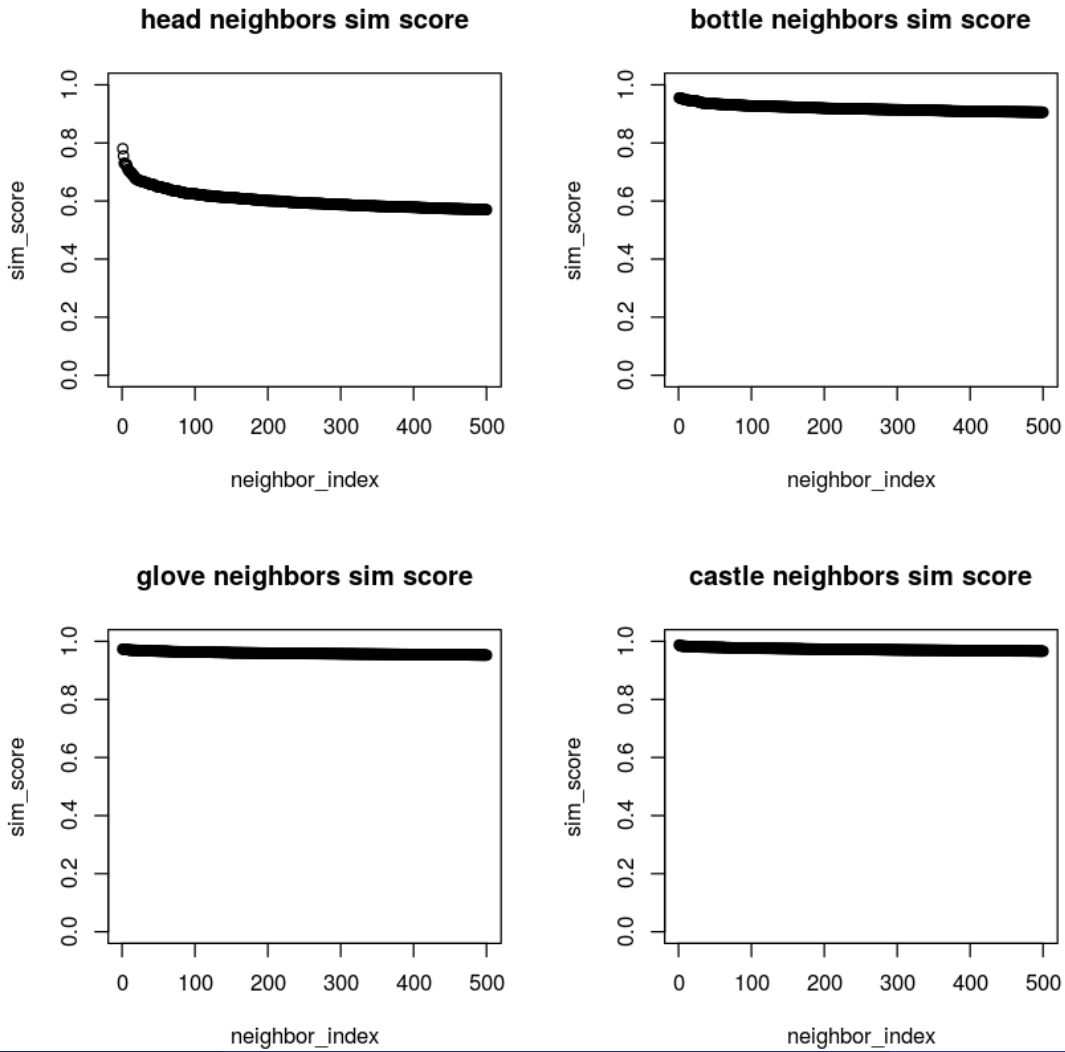
# Evaluation à l'aide de BLESS : ACL w2vf



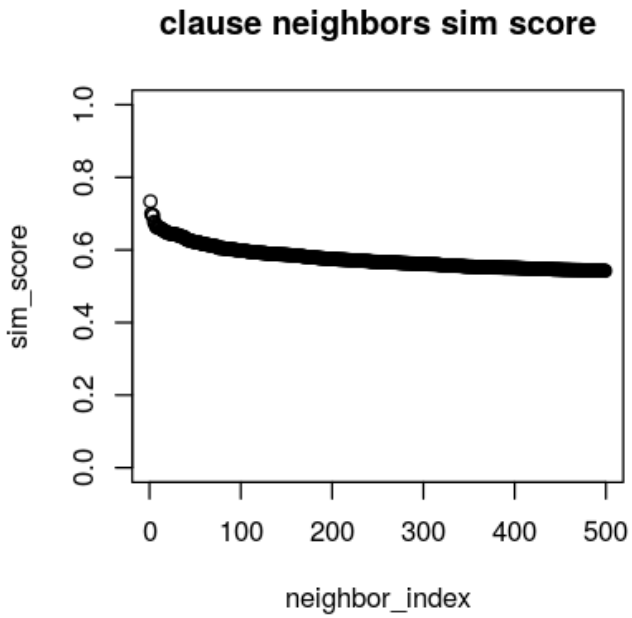
# Embeddings générés avec w2vf

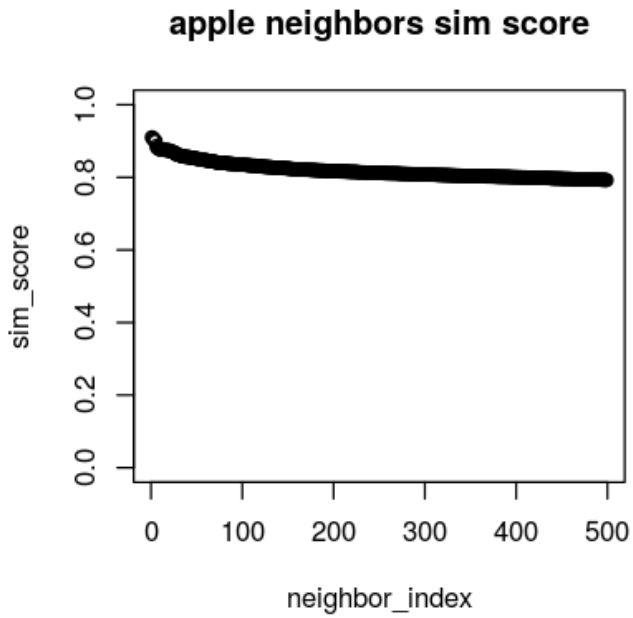
- Que se passe t-il avec la relation random pour les noms ?
- Sélection de mots spécifiques au corpus et d'autres moins spécifiques
- Observation du score de similarité obtenu pour les 500 premiers voisins de ces mots

# Observation du score de similarité des 500 premiers voisins pour mots choisis

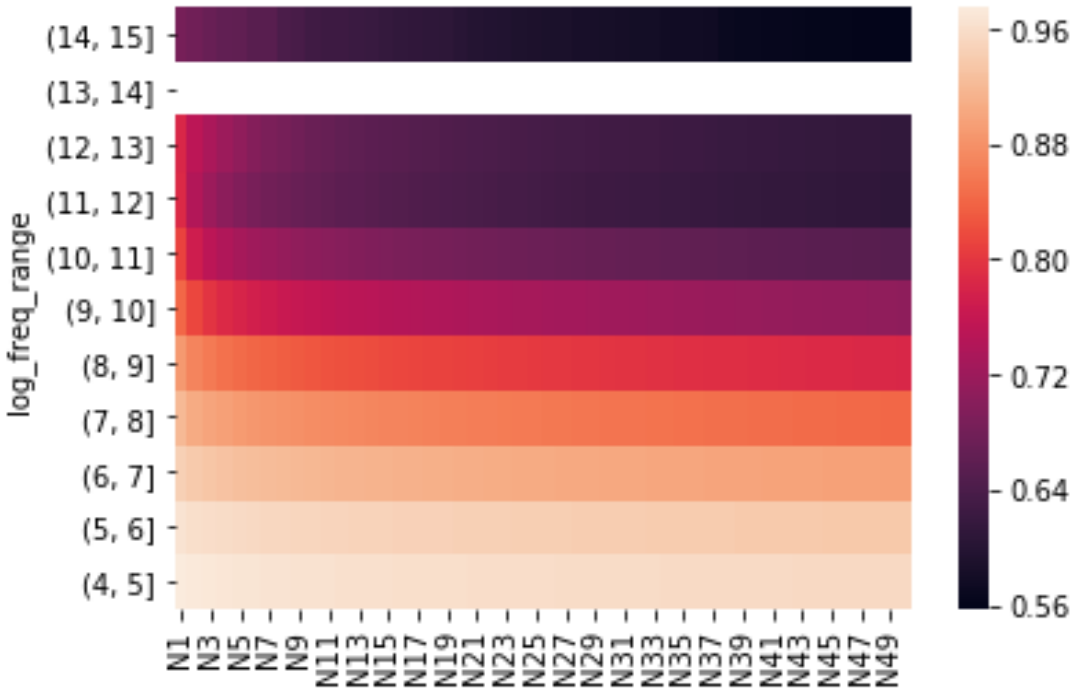


# Observation du score de similarité des 500 premiers voisins pour mots choisis



# Embeddings générés avec w2vf : rôle de la fréquence ?




Est-ce qu'on a affaire à des zones dans lesquelles le modèle n'arrive pas à distinguer le sens des mots ?

# Plan

# Perspectives

- Approfondir les observations faites sur les adjectifs
- Question de la fréquence notamment avec word2vecf
- Variation sur plus de paramètres
  - Taille des vecteurs
  - Filtrage mots en dessous seuil fixé
  - Taille fenêtre contexte
  - Triplets syntaxiques
- Quelle influence des paramètres sur les différents modèles générés ?
- Evaluation à l'aide d'un lexique d'évaluation

# Références


Baroni, M. and Lenci, A. (2011). How we BLESSed distributional semantic evaluation. *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*.

Bernier-Colborne, G. and Drouin, P. (2016). Evaluation of distributional semantic models : a holistic approach. *Proceedings of the 5th International Workshop on Computational Terminology*, pages 52–61.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Fan Tan, Y. (2008). The ACL Anthology Reference Corpus : A Reference Dataset for Bibliographic Research in Computational Linguistics. *Proceedings of Language Resources and Evaluation Conference (LREC 08)*.

Bowman, S., Goldberg, Y., Hill, F., Lazaridou, A., Levy, O., Reichart, R., and Sogaard, A. (2017). RepEval 2017 : The Second Workshop on Evaluating Vector Space Representations for NLP.

Colyer, A. (2017). The amazing power of word vectors.

Consortium, B. (2007). The British National Corpus, version 3 (BNC XML Edition).

Fabre, C. and Lenci, A. (2015). Distributional Semantics Today Introduction to the special issue. *TAL*, 56(2) :7–20.

Faruqui, M. and Dyer, C. (2014). Community Evaluation and Exchange of Word Vectors at wordvectors.org. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 19–24.

Faruqui, M., Tsvetkov, Y., Rastogi, P., and Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. *arXiv :1605.02276*.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing Search in Context : The Concept Revisited. *ACM Transactions on Information Systems*, 20(1) :116 :131–116 :131.

Gavagai (2015). A brief history of word embeddings (and some clarifications).

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., and Weese, J. (2013). UMBC EBIQUITY-CORE : Semantic Textual Similarity Systems. *Proc. 2nd Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics*.

Jurafsky, D. and Martin, J. H. (2017). Semantics with Dense Vectors. In *Speech and Language Processing. Draft of November 7, 2016*.

Levy, O. and Goldberg, Y. (2014). Dependency-Based Word Embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 302–308.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv :1301.3781*.

Nayak, N., Angeli, G., and Manning, C. D. (2016). Evaluating Word Embeddings Using a Representative Suite of Practical Tasks. *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (September) :298–307.

Tanguy, L., Sajous, F., and Hathout, N. (2015). Evaluation sur mesure de modèles distributionnels sur un corpus spécialisé :