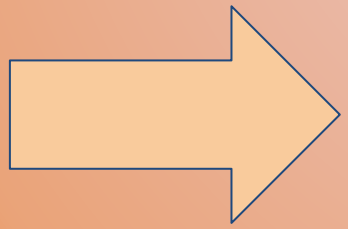
Projet ADDICTE

Analyse distributionnelle en domaine de spécialité

Corpus de domaine spécialisé

Ensemble de caractéristiques :

- Structures de documents spécifiques (titres, introduction, méthodologie, légendes…)
- Éléments particuliers (listes, titres de sections)
- Lexique et termes techniques complexes



Impact sur ADA

T3.1 : Caractéristiques linguistiques des contextes distributionnels

Objectifs :

1) **Constitution d'un corpus en domaine de spécialité (TAL)** structuré selon la norme TEI et diffusable sur ORTOLANG.
2) **Étude des caractéristiques linguistiques des contextes** dans des parties spécifiques de document.

Quel est leur impact ? Peut-on les catégoriser ?

1) Constitution du Corpus TALN

- Articles scientifiques des conférences TALN et RECITAL des années 1997 à 2019.
- **Travaux antérieurs :** Constitution des archives PDF par Florian Boudin avec métadonnées des articles. Extraction du texte brut par Ludovic Tanguy pour un premier corpus non structuré au format TXT.

Total : 1579 articles - 5.5 Mmots

Éléments à extraire et annoter.

Résumé/abstract
Keywords

Titres de sections classés par :
- niveau
- fonction récurrente (intro/méthodo/conc)

Structure PDF d'un article de recherche TALN

Yoann BARD, Laboratoire CLLE - Maison de la Recherche
5 Allée Antonio Machado, 31058 Toulouse cedex 9.
yoann.bard@univ-tlse2.fr

Résumé - Abstract.
Ceci est le contenu textuel d'un résumé à extraire dans le traitement.
This is the text of the abstract to extract in the process.

Mots Clés - Keywords.
Fouille de texte, extraction, corpus spécialisé, conversion PDF, analyse sémantique distributionnelle.

1 Introduction.

Les actes des conférences RECITAL/TALN sont tous au format PDF. Le ***Portable Document Format***, communément abrégé en **PDF**, est un langage de description de page présenté par la société Adobe Systems en 1992 et qui est devenu une norme ISO en 2008. La spécificité du PDF est de préserver la mise en page d'un document – polices de caractère, images, objets graphiques, etc. – telle qu'elle a été définie par son auteur, et cela quels que soient le logiciel, le système d'exploitation et l'ordinateur utilisés pour l'imprimer ou le visualiser.

2 Définition.

2.1 Contraintes

Le format PDF est conçu de manière graphique pour l'oeil humain et donc très compliqué à traiter dans un processus informatique. Il faut différencier les éléments tels que les titres, contenu, tableaux, figures, header/footnotes et références. Il existe cependant certains patterns graphiques utilisées et il reste possible de différencier selon leur position, fontsize et parfois leur contenu.

2.2 Différents éléments

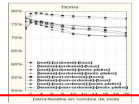


Figure 1 : explications sur la figure

Table 1 : Matrice de confusion.

3 Conclusion

Nous avons vu les éléments principaux qui composent un article de recherche TAL classique. Bien sûr ils peuvent plus ou moins varier de cette structure. (il y a aussi la bibliographie qui serait à extraire).[1]

[1] Ceci est une footnote à extraire

(+bibliographie)

parasites (n°page, headnotes, titres…)

Métadonnées article : titre, auteur, date...

Contenu textuel de section
(peut contenir des références et des formules)

Légendes tableaux/ figures.

Footnotes

Outils de conversion PDF (opensource)

- **PDFMiner :** Library python (2018)
- **pdf2xml** : exécutable (2016)
- **pymupdf :** library python (2019)

Permet une conversion en HTML/TXT/XML respectant au mieux la structure graphique du document.
Segmentation générique des éléments au caractère/mot ou ligne par ligne (peut varier selon les options des outils).

Informations graphiques sur chaque élément : position x/y, fontsize, fontname, largeur, hauteur, bold, italic...

Exemple Output HTML :

<p style="X:109;Y:79;">contenu textuel de la ligne

1 : Conversion PDF2XML + ParsCit

PDF file	
PDF extraction	`xml: content & layout`  `(pdf2xml/Xpdf)`
Segmentation model	CRF Segmentation
Segmented document	cover · header · body · foot-notes · head-notes · biblio · Annexes

2 : Cleaning semi-automatique et restructuration des documents

Balisage des résumés

- Erreurs de balisage
- Notes de bas de pages
- Numéros

- Balisage des figures

- Équations

Parties de textes en anglais.

Césures

```
RESUME
</figure>
<bodyText confidence="0.997599588235294">
Nous décrivons notre méthode structurée en
trois étapes : construction et pré-traitement du réseau, détection de la structure de communautés,
construction des plongements de mots à partir de cette structure. Après avoir décrit cette nouvelle
méthodologie, nous montrons la pertinence de notre approche avec des premiers résultats d'évaluation
sur les tâches de catégorisation et de similarité. Enfin, nous discutons des perspectives importantes
d'un tel modèle issu des réseaux complexes : les dimensions du modèle (les communautés) semblent
interprétables, l'apprentissage est rapide, la construction d'un nouveau plongement est presque
instantanée, et il est envisageable d'en expérimenter une version incrémentale pour travailler sur
des
corpus textuels temporels.
</bodyText>
<sectionHeader confidence="0.943186" genericHeader="abstract">
ABSTRACT
</sectionHeader>
<bodyText confidence="0.9856198571428571">
Complex networks based word embeddings.
Most of the time, the first step to learn word embeddings is to build a word co-occurrence matrix.
Community structure is used as a way to
reduce the dimensionality of the initial space. Using this community structure, we propose a method
to extract word embeddings that are comparable to the state-of-the-art approaches.
</bodyText>
<figure confidence="0.709760125">
MOTS-CLÉS : Plongements lexicaux, réseaux complexes, détection de communautés.
KEYWORDS: Word embeddings, complex networks, community detection.
Apprentissage de plongements lexicaux par une approche réseaux complexes
TALN-RECITAL@PFIA 2019
28
Articles longs
1
Introduction
</figure>
<bodyText confidence="0.992479260292683">
Dans l'état de l'art de l'apprentissage de plongements lexicaux, on recense de nombreuses approches
basées sur une matrice de co-occurrences termes-termes construite en utilisant de grands corpus
(Pennington et al., 2014; Levy et al., 2015). Les auteurs factorisent ensuite cette matrice creuse
de façon à
obtenir un nouvel espace dans lequel chaque terme est représenté par un vecteur dense.
</bodyText>
<bodyText confidence="0.882915888888889">
TALN-RECITAL@PFIA 2019
FIGURE 1 – Distribution de la taille des communautés (en log-log)
Par ailleurs, la taille des communautés a une influence similaire. L'algorithme de détection de com-
munautés aboutit (sauf exception) à une partition dont les tailles des communautés sont hétérogènes.
</bodyText>
<equation confidence="0.96807275">
n =
n − µ(ˆec
ˆec
```

Cleaning automatique

Mauvais balisage

Langues étrangères

Métadonnées

<equation confidence="0.479462">
{prØnom.nom}@paris4.sorbonne.fr
RØsumØ (cid:150) Abstract
</equation>
<bodyText confidence="0.968887071428571">
Cet article prØsente tout d(cid:146)abord une analyse linguistique des cad
implØmentation informatique. Puis (cid:224) partir de ce travail, une mod
(cid:224)
l(cid:146)ensemble des cadres de discours est proposØe. Enfin, nous discu
d(cid:146)indicateur
proposØ dans le cadre thØorique de l(cid:146)exploration contextuelle.
To begin with, this paper outlines a linguistic analysis of textual enumerat
its computational making. Then, from this work, a modelling for all textua
suggested. Finally, we discuss the relevance of the concept of clue which i
theoretical framework of contextual exploration method.
Keywords (cid:150) Mots ClØs
SØries de cadres organisationnels, marqueurs d(cid:146)intØgration linØa
segmentation automatique de textes, mØthode d(cid:146)exploration cont
de textes.
Enumerating frameworks, linear integration markers, discourse frames, au
segmentation, contextual exploration method, automatic text filtering.
سايساعمتجم et un autre civil
</bodyText>
<footnote confidence="0.989235285714286">
3. http://www.who.int/mediacentre/factsheets/fs312/fr/
4. http://www.who.int/mediacentre/factsheets/fs138/fr/
5. Integrating Informatics and Biology to the Bedside, https://www.i2l
[1]
</footnote>

Encodage de caractères spéciaux:
^e -> ê
cid:224 -> à
Ø -> é

URL

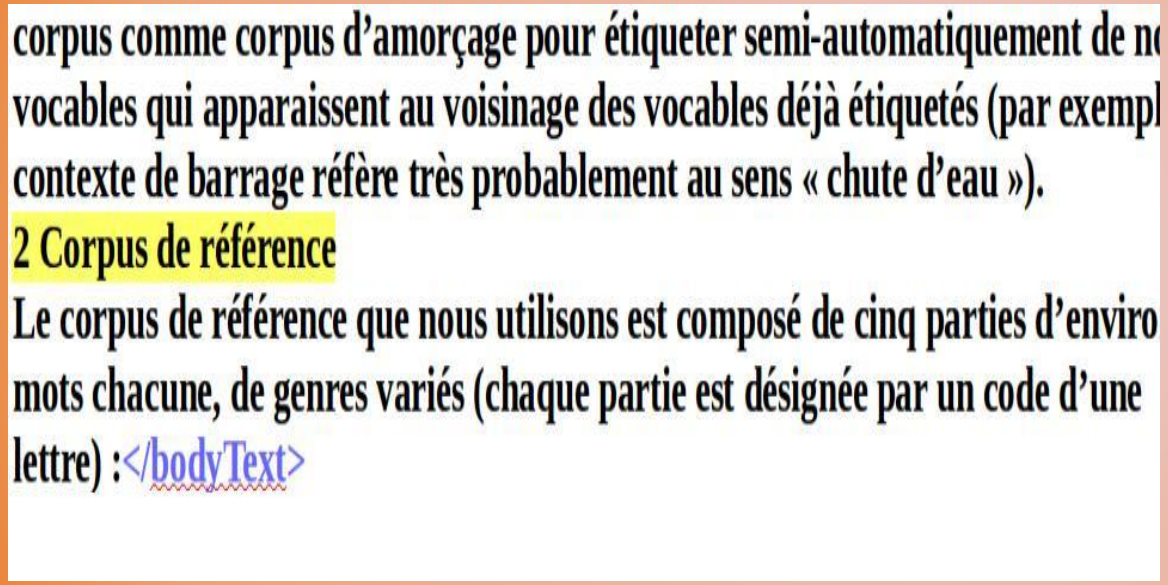
Cleaning automatique

A l’opposé, le verbe entrer possède
nécessairement le sous graphe relationÈpartieÈfinÈtopologie et exclut la partie relationÈtout. (?? entrer autour de...)
3.3.2. contraintes et relations curvilignes :</bodyText>
<bodyText confidence="0.956662958333333">Nous cherchons en particulier à mett
règles générales d’exclusion : passer par tel noeud du graphe impose telle restrictio
ii Il s’agit ici, et plus bas, du OU EXCLUSIF.
Yann Mathet
la suite du parcours (par ex. une contrainte de forme ne peut concerner le domaine topologique).
la suite du parcours (par ex. une contrainte de forme ne peut concerner le domaine topologique).
Le but est de montrer comment s’articulent les parties de sens que
nous avons mises au jour (en particulier les relations de forme ou qui existaient dan
d’autres travaux (comme la topologie).
5. Conclusion
Cette vision, plutôt que de s’opposer à celles que nous avons étudiées, tente de les co
dans un graphe du domaine qui fonde ses relations sur tous les objets du domaine, s

A l’opposé, le verbe entrer possède
nécessairement le sous graphe relationÈpartieÈfinÈtopologie et exclut la partie relationÈtout. (?? entrer autour de...)</bodyText>
<subsubsectionHeader n='3.3.2'>contraintes et relations curvilignes :</subsubsectionHeader>
<bodyText confidence="0.956662958333333">Nous travaillons actuellement à la délimitation
parties du graphe qui sont concernées
par chacun des verbes de déplacement. Nous cherchons en particulier à mettre au jour des règles générales d’exclusion : passer par tel noeud du graphe impose telle restriction quant à la suite du parcours (par ex. une contrainte de forme ne peut concerner le domaine topologique).
Le but est de montrer comment s’articulent les parties de sens que
nous avons mises au jour (en particulier les relations de forme ou qui existaient dans d’autres travaux (comme la topologie).</bodyText>
<sectionHeader id='conclusion'>Conclusion</sectionHeader>
<bodyText>
Cette vision, plutôt que de s’opposer à celles que nous avons étudiées, tente de les concilier dans un graphe du domaine qui fonde ses relations sur tous les objets du domaine, sans

Cleaning semi-automatique

Pour les corrections plus ambigües :


corpus comme corpus d'amorçage pour étiqueter semi-automatiquement de n
vocables qui apparaissent au voisinage des vocables déjà étiquetés (par exempl
contexte de barrage réfère très probablement au sens « chute d'eau »).
2 Corpus de référence
Le corpus de référence que nous utilisons est composé de cinq parties d'enviro
mots chacune, de genres variés (chaque partie est désignée par un code d'une
lettre) :</bodyText>

Deux approches :

Diagnostique avec liste de titres extraits des documents HTML.
Si match == **correction**.
Sinon :

Ajout de commentaire XML pour **vérification manuelle**.

Autres exemples :

Footnotes :

```
<footnote>
 2 Nom masculin singulier défini au génitif. nom féminin singulier construit au nominatif.
</footnote>

<footnote confidence="0.451125">
2 Analyseurs syntaxiques robustes
Les outils d'analyse syntaxique robustes sont des systèmes conçus pour pouvoir marquer
</footnote>
```

Légendes :

```
13 867
82 730
90 248
TABLE 2 – Sélecteurs initiaux sur l'ensemble des traductions (noms, verbes et adjectifs). La
</figure>
<bodyText confidence="0.944481">
couverture C est le nombre total de paires (littéral, synset).
La table 2 montre les résultats de cette opération. La couverture donne une idée de la taille
```

Césures

corpus comme corpus d'amorçage pour étiqueter semi-automatiquement de nouvelles formes qui apparaissent au voisinage des mots-clés déjà étiquetés (par exemple chute dans le con-texte de barrage réfère très probablement au sens « chute d'eau »).

Création d'un lexique de mots-composés du corpus TALN.

Vérification

semi-automatique ✓

con-texte ✗ → **contexte**

Métadonnées

Titres (sous-)sections

Légendes et footnotes

Contenu textuel

Bibliographie

```
<resume>
Dans le cadre de la recherche en sémantique lexicale, l'équipe TAL du LIRMM développe actuellement
un système d'analyse des aspects thématiques des textes et de désambiguisation lexicale basé sur le
vecteurs conceptuels. Pour la construction des vecteurs, les définitions provenant de sources
lexicales différentes (dictionnaires à usage humain, listes de synonymes, définitions de thésaurus,
) sont analysées.
</resume>
<mots_cles>
Traitement automatique des langues naturelles, classification automatique, désambiguïsation
sémantique lexicale
</mots_cles>
<title/>
<abstract>
In the framework of research in meaning representation in NLP, we focus our attention on thematic
aspects and conceptual vectors. A vectorial base is built upon a morphosyntactic analysis of severa
lexical resources to reduce isolated problems. A conceptual vector is associated with each
definition and another one with the global meaning of a word.
</abstract>
<keywords>
Natural language processing, unsupervised clustering, word sense disambiguation
</keywords>
<body>
<sectionHeader id="introduction">
Introduction
</sectionHeader>
<bodyText confidence="0.966065558823529">
Dans le cadre de la recherche en sémantique lexicale, l'équipe TAL du LIRMM développe actuellement
un système d'analyse des aspects thématiques des textes et de désambiguisation lexicale basé sur le
vecteurs conceptuels. Pour la construction des vecteurs, les définitions provenant de sources
lexicales différentes (dictionnaires à usage humain, listes de synonymes, définitions de thésaurus,
) sont analysées.
</bodyText>
<sectionHeader confidence="0.524714" n="1">
 Catégorisation des définitions
</sectionHeader>
<bodyText confidence="0.996288692307692">
Les méthodes de classification automatique sont nombreuses (Alpert, Kahng, 1995 (Berkhin, 2002) mai
ne sont cependant pas directement applicables dans le cadre de cette étude, car la catégorisation e
la classification dans ces domaines traitent un grand nombre de données à répartir dans un faible
nombre de classes en un processus unique. Dans notre cas, la masse
de données est importante (actuellement pour 110 000 termes, plus de 430 000 définitions et vecteur
conceptuels mais la catégorisation porte sur les définitions d'un seul terme (environ 5 définitions
en moyennes par source, certains termes fortement polysémiques peuvent en avoir plus de 50).
</bodyText>
<subsectionHeader confidence="0.996599" n="1.1">
Choix de l'algorithme
</subsectionHeader>
<bodyText confidence="0.987754916666667">
Le choix de l'algorithme repose de différents constats :
Le volume de la donnée est faible. Le problème est donc moins restrictif concernant le choix
des méthodes d'analyses et de l'algorithme. Cependant, il est impossible d'envisager un
entrainement sur lequel repose certains algorithmes dont les Support Vector Machine
(SVM) par exemple (Vapnik, Chervonenkis, 1964 (Burges, 1998).
<bodyText>
<figureCaption confidence="0.99939" type="figure" n="1">
Comportement d'un système fermé avec 20 agents et 20 objets
</figureCaption>
<reference confidence="0.9823863125">
C.J. Alpert, A.B. Kahng Recent Directions in Netlist Partitioning : A Survey Integration : VLSI
J., vol. 19, 1995, 93 pp.
</reference>
```

État actuel du Corpus TALN

- 1579 articles des conférences TALN et RECITAL des années 1997 à 2019.
- Fichiers au format XML structurés par des balises séparant les différentes parties (métadonnées,titres…)
- Nombre total de mots ≈ **5,5Mmot**
  Corpus ACL = 90 Mmots
  Corpus PMC OA = 2 M articles

A venir : correction par IE (contrat CORLI) et diffusion sur ORTOLANG.

2) Étude des contextes distributionnels

Outils

Méthodologie

"Phrases du corpus"

TALISMANE

Parsing et lemmatisation du Corpus

N:phrase D:du N:corpus

Python

Matrice de cooccurrences graphiques

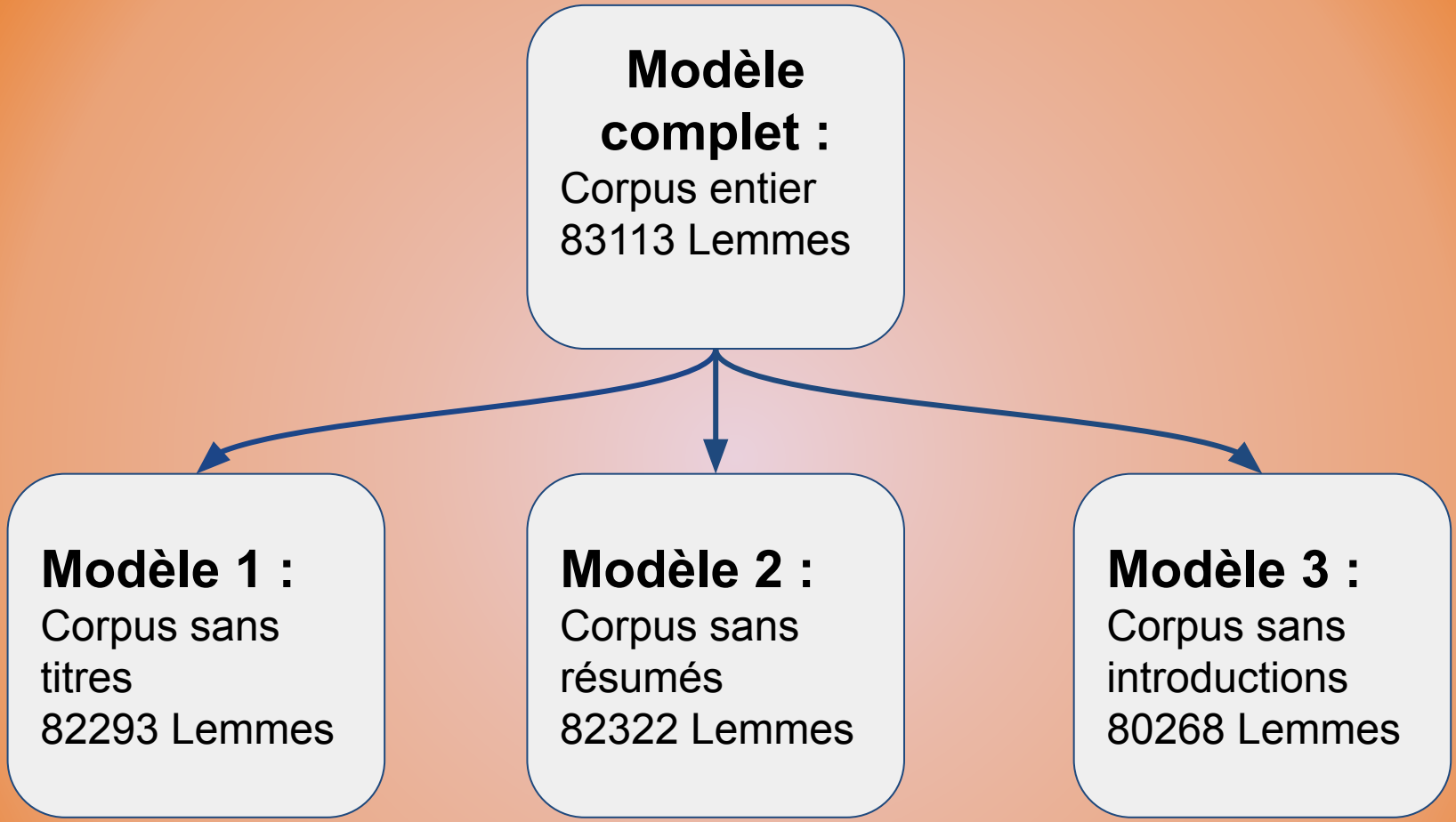
N:phrase N:corpus 1

DISSECT
M.Baroni
G.Dini
N. Pham

Modèle distributionnels (SV)

N:phrase N:corpus PPMI

Script d'extraction de contextes.


Modèle complet :
Corpus entier
83113 Lemmes
Modèle 1 :
Corpus sans titres
82293 Lemmes
Modèle 2 :
Corpus sans résumés
82322 Lemmes
Modèle 3 :
Corpus sans introductions
80268 Lemmes

N:langage

Voisins modèle complet

1	N:grammaire
2	ADJ:naturel
3	N:langue
4	N:modèle
5	ADJ:linguistique
6	N:formalisme
7	N:automate
8	N:représentation
9	ADJ:formel
10	N:modélisation

Voisins modèle sans titres

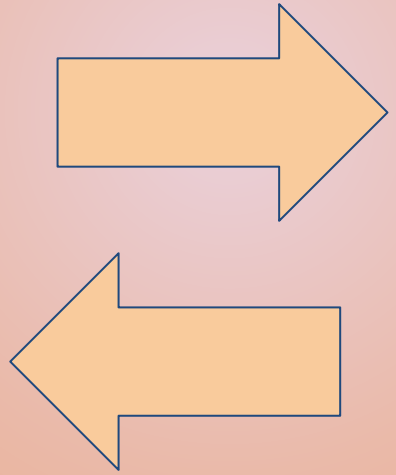
N:grammaire
ADJ:naturel
N:modèle
N:langue
N:formalisme
N:représentation
ADJ:linguistique
N:automate
ADJ:formel
N:modélisation

Analyse des paires

Première étape :

Comparaison des paires de voisins avec la mesure de similarité RBO (W.WEBBER A.MOFFAT J.ZOBE)

Exemple :
Voisins du mot LANGAGE entre MODÈLE 1 et 2



RBO = 0.767

RBO = 1 --> Mot stable
RBO = 0 --> Mot instable

1	N:grammaire	N:grammaire
2	ADJ:naturel	ADJ:naturel
3	N:langue	N:modèle
4	N:modèle	N:langue
5	ADJ:linguistique	N:formalisme
6	N:formalisme	N:représentation
7	N:automate	ADJ:linguistique
8	N:représentation	N:automate
9	ADJ:formel	ADJ:formel
10	N:modélisation	N:modélisation

V:assister

Voisins modèle complet

1	N:traducteur
2	N:outil
3	V:destiner
4	N:aide
5	V:aider
6	ADJ:humain
7	N:expert
8	ADJ:automatique
9	V:développer
10	N:ordinateur

N:aide
ADJ:automatique
N:outil
N:programme
N:traducteur
N:ordinateur
N:expert
V:destiner
ADJ:humain
N:traduction

Voisins modèle sans résumés

RBO = 0.1

V:assister + N:traducteur

Modèle complet

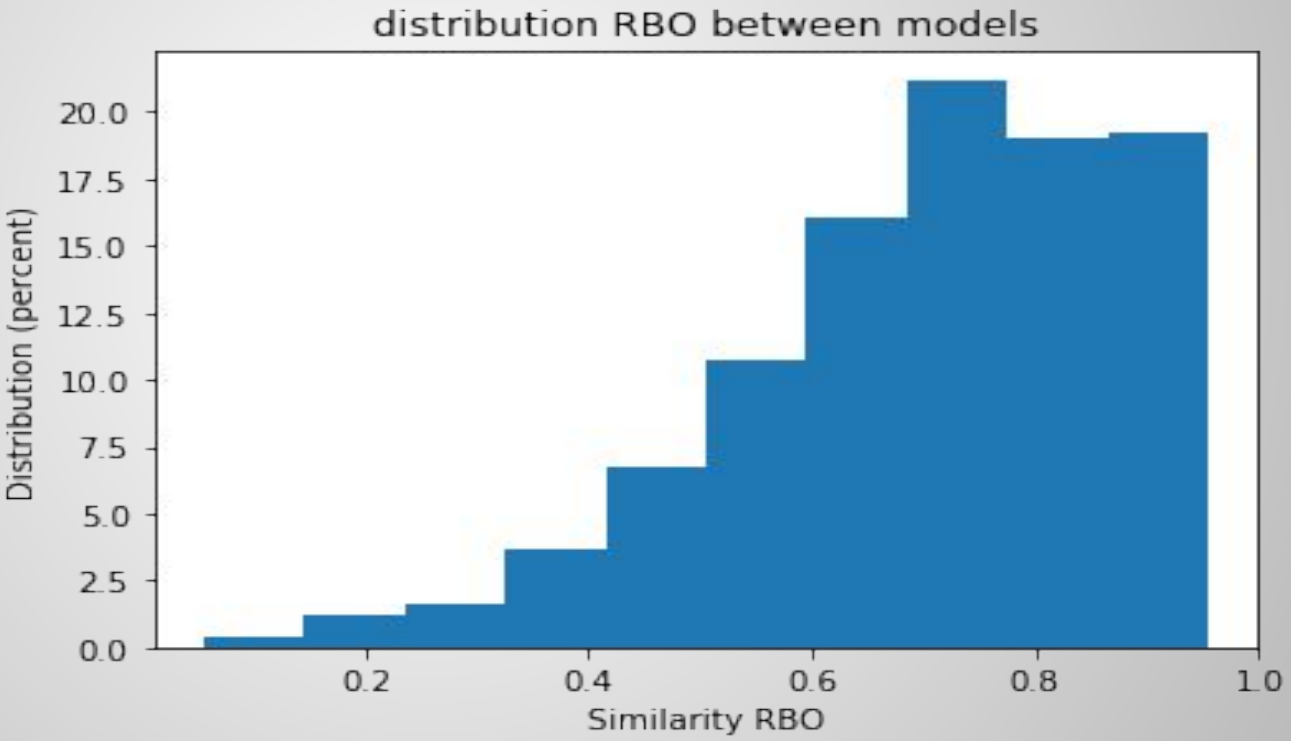
Co-contextes	Score	Co-contextes	Score
N:rédacteur	0.02005	N:lexicographe	0.01871
N:lexicographe	0.01843	ADJ:humain	0.01371
N:préparation	0.01720	N:usager	0.01133
N:usager	0.01361	PROPN:multi-lingue	0.01077
ADJ:humain	0.01322	N:interrogation	0.01062
ADJ:littéraire	0.01085	ADJ:littéraire	0.01036
N:interrogation	0.00946	N:traducteur	0.01023
PROPN:multi-lingue	0.00908	V:aider	0.00951
N:traducteur	0.00894	N:mémoire	0.00936

Modèle sans résumés

Distribution RBO
Modèle 1 (entier) vs 2 (sans les titres)

Distribution sur **3040** mots communs avec fréquence Min = 50

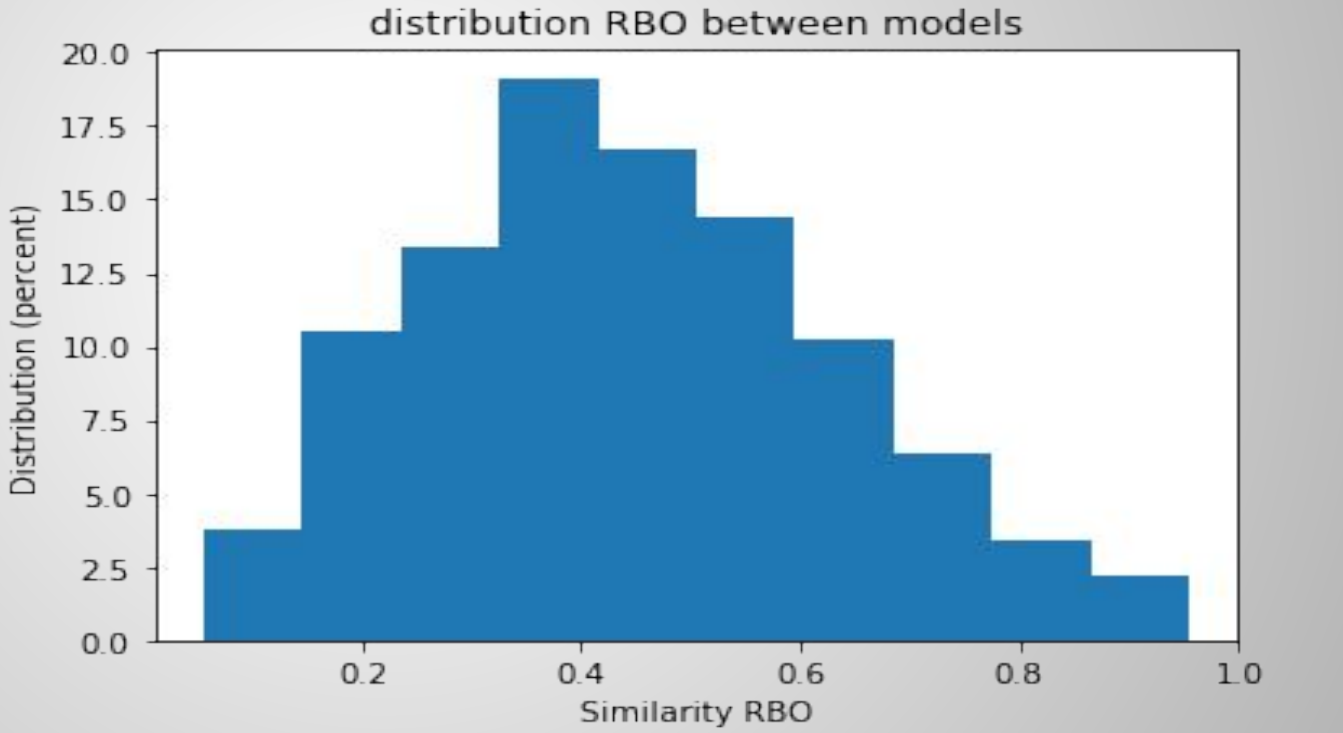
RBO moyen = 0.745



Distribution RBO
Modèle 1 (entier) vs 3 (sans les résumés)

Distribution sur **3011** mots communs avec fréquence Min = 50

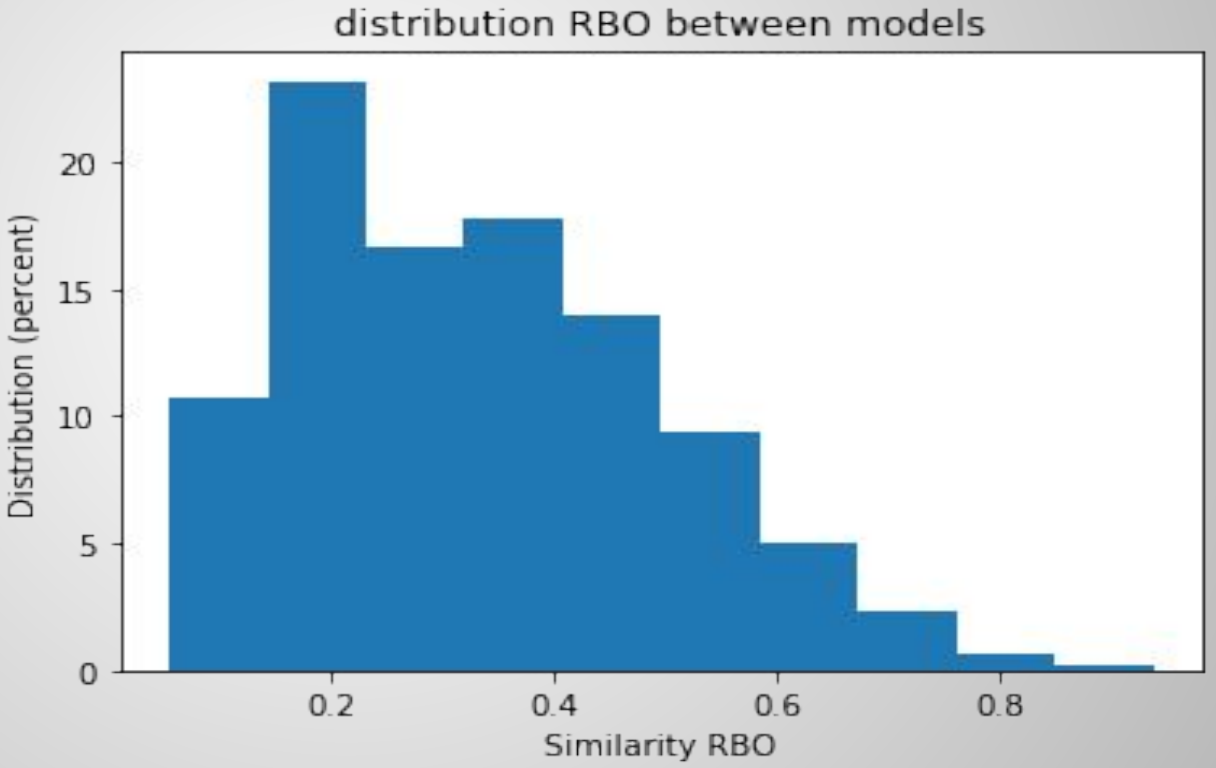
RBO moyen = 0.496



Distribution RBO
Modèle 1 (entier) vs 4 (sans les introductions)

Distribution sur **2947** mots communs avec fréquence Min = 50

RBO moyen = 0.377



Observations

1. Les mots avec une forte baisse de fréquence sont souvent instables.

Modèle entier vs Modèle sans titres :

conclusion : **Freq1** = 1442 **Freq2** = 430 **RBO** = 0.19

introduction : **Freq1** = 1505 **Freq2** = 353 **RBO** = 0.29

discussion : **Freq1** = 495 **Freq2** = 276 **RBO** = 0.26

2. De nombreux mots avec des voisins peu similaires sont instables (peu importe la fréquence). Exemples :

ADJ:basique
RBO = 0.1
Perte freq = 77-67

1	ADJ:évolutif	V:renforcer
2	V:renforcer	N:exception
3	ADJ:abstraire	ADJ:analogique
4	N:exception	ADJ:porteur
5	ADJ:analogique	ADJ:abstraire

V:orienter
RBO=0.1
437-394

1	V:faciliter	N:interaction
2	ADJ:automatique	ADJ:actuel
3	N:développement	ADJ:automatique
4	N:interaction	N:développement
5	V:viser	V:mener

N:encodage
RBO = 0.1
116-106

1	V:découler	ADJ:polylexicales
2	N:treillis	ADJ:composite
3	V:encoder	V:encoder
4	ADJ:composite	V:découler
5	ADJ:polylexicales	ADV:parfois

3. Les mots avec une faible baisse de fréquence peuvent être instables. Exemple modèle sans résumé :

N:paramétrage
RBO = 0.1
Perte freq = 62-60

1	N:ajustement	N:paramètre
2	N:paramètre	V:rester
3	N:adaptation	N:soin
4	V:implémenter	N:ajustement
5	N:résultat	V:renforcer

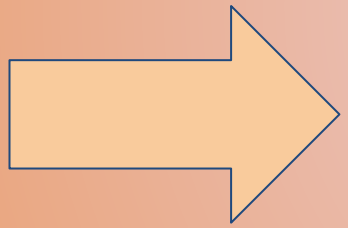
ADJ:marginal
RBO=0.1
70-68

1	ADJ:notable	ADJ:notable
2	ADJ:significatif	V:moyenner
3	V:dégager	ADJ:significatif
4	V:moyenner	V:rapporter
5	V:rapporter	N:productivité

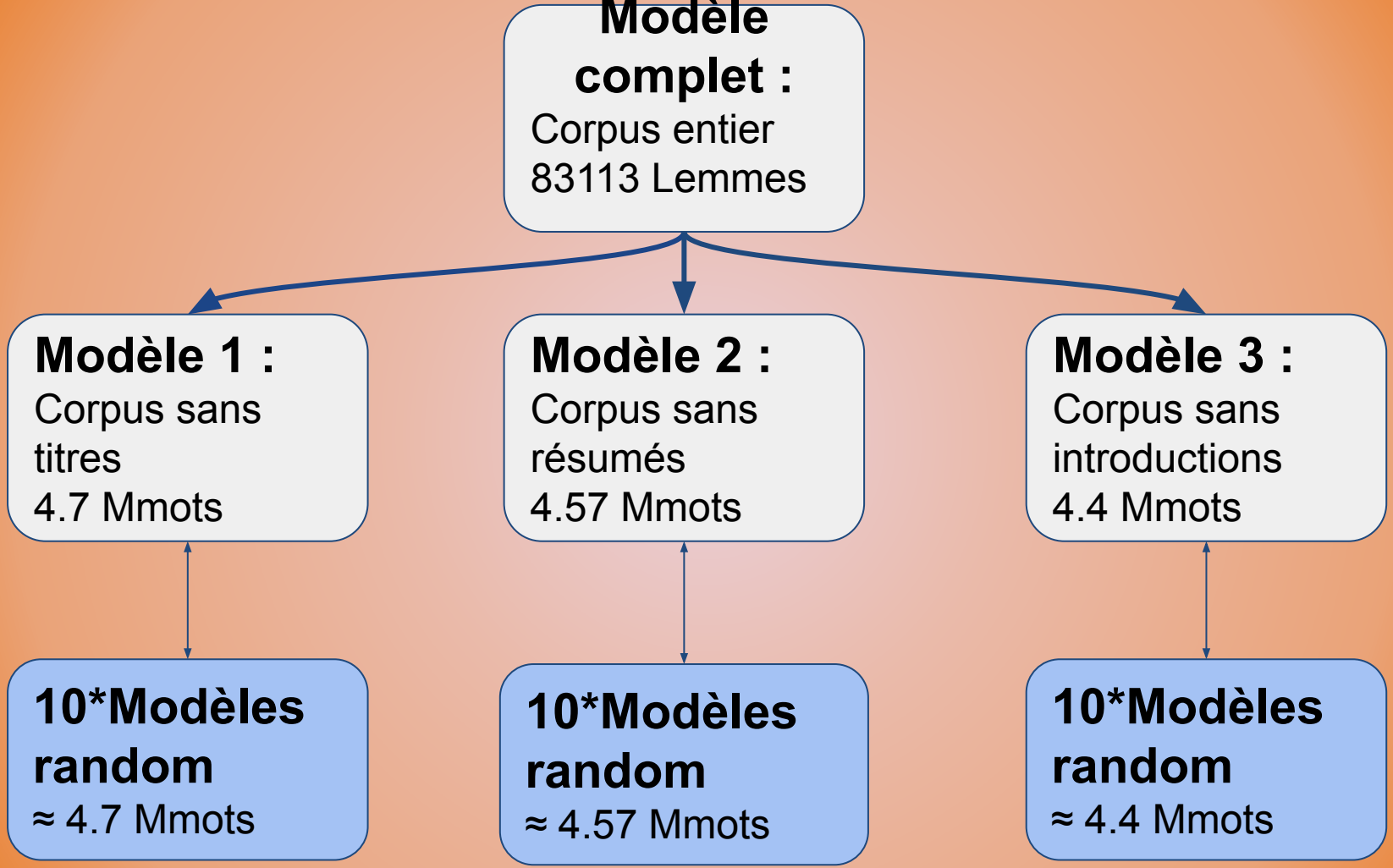
ADV:grammaticalement
RBO = 0.1
55-55

1	ADV:sémantiquement	ADV:syntaxiquement
2	ADV:syntaxiquement	ADJ:cohérent
3	ADJ:correct	ADJ:correct
4	V:être	ADV:linguistiquement
5	ADJ:cohérent	V:être

L’instabilité est-elle causée seulement par la perte de mots ou bien la structure du document étudiée a aussi un impact important ?



Création de modèles aléatoires


Modèle complet :
Corpus entier
83113 Lemmes
Modèle 1 :
Corpus sans titres
4.7 Mmots
Modèle 2 :
Corpus sans résumés
4.57 Mmots
Modèle 3 :
Corpus sans introductions
4.4 Mmots
10*Modèles random
≈ 4.7 Mmots
10*Modèles random
≈ 4.57 Mmots
10*Modèles random
≈ 4.4 Mmots

Modèle 1 :
Corpus **sans titres**
4.7 Mmots

RBO moy = 0.745

VS

RBO moy = 0.729

10*Modèles random
≈ 4.7 Mmots

Modèle 2 :
Corpus **sans résumés**
4.57 Mmots

RBO moy = 0.489

VS

RBO moy = 0.468

10*Modèles random
≈ 4.57 Mmots

Modèle 3 :
Corpus **sans introductions**
4.4 Mmots

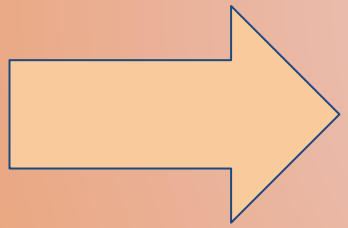
RBO moy = 0.377

VS

RBO moy = 0.378

10*Modèles random
≈ 4.4 Mmots

Les structures de document étudiées ont-elles un impact sur la distribution des mots spécifiques du domaine spécialisé TAL ?



Calcul de spécificité Chi2 + Log Evert

Observations : 4 cas de figures

Spec | Freq	Stabilité -	Stabilité +
+ +	Mots subissant une **grosse perte de fréquence** : démonstration assisté veille arborer	Mots présents dans des **zones stables :** - Verbes d'exposition (présenter, aborder, finaliser) - Évaluatifs (vaste, important, faiblement)
- -	Mots **polysémiques ou dispersés** : - synthétique (voix/ aperçu) - interrompre (parole, locuteur/ action) - équation	Mots présents dans des **zones stables :** - possessif - attributif - pattern - noeud - sous-arbre

	Corrélation Chi2	Corrélation Evert
RBO1 (sans titres)	0.0273	-0.0570
RBO2 (sans résumés)	0.0838	0.1540
RBO3 (sans intros)	0.1021	0.2107

Résultats : Plus le Chi2 / Log Evert est élevé et plus le RBO est stable (Coef Pearson positif).

Intuition : Les mots spécifiques sont moins affectés par l'absence des résumés ou des intros.

Conclusion et perspectives

- Le coeur du vocabulaire technique n'est pas présent dans les résumés, intros ou titres de sections.
- Ces structures n'ont pas beaucoup d'impact sur l'analyse distributionnelle en général.
- Processus d'évaluation intéressant, mise en place de nouvelles mesures pour étudier les variations de voisins/contextes.
- Perspective : étudier les contextes syntaxiques.
- Perspective plus importante : nettoyage plus poussé du corpus TALN pour avoir des modèles propres lors des prochaines analyses.