Référence de la version finale :
Cocco C., Catégorisation automatique de propositions textuelles en types de discours. In : Clivaz C., Meizoz J., Vallotton F. & Verheyden J. (Eds.) *Lire demain : des manuscrits antiques à l'ère digitale = Reading tomorrow : from ancient manuscripts to the digital era.* Presses polytechniques et universitaires romandes, Lausanne, pp. 689-707, 2012.

Catégorisation automatique de propositions textuelles en types de discours

Christelle Cocco

Résumé
De nombreuses méthodes de classification automatique de textes (supervisée ou non) existent. Elles se sont particulièrement développées depuis la création du web en vue d'améliorer les moteurs de recherche. Au delà de ce champs d'application, elles peuvent aussi être utiles pour « découvrir » ou « redécouvrir » ce qui caractérise une classe de textes aussi simplement que possible. Dans l'approche statistique développée ici, il s'agira de partitionner des parties de textes en types de discours tels que narratif, explicatif, argumentatif etc. en utilisant un minimum d'information, à savoir les catégories morphosyntaxiques (CMS). Les premiers résultats obtenus mettent en évidence des liens significatifs entre les CMS, les types de discours et des textes annotés par un expert humain ; ces relations sont toutefois encore incomplètes et difficiles à interpréter d'un point de vue linguistique.

Abstract
Numerous automatic classification or clustering methods of texts exist. They have been particularly developed since the Web creation to improve search engines. Beyond that, they can also be useful to "discover" and to "rediscover" what characterises, as simple as possible, a class of texts. In the statistical approach developed here, the aim is to cluster text parts into discourse types, such as narrative, explicative, argumentative, etc. with minimal information, namely the part-of-speech (POS) tags. The preliminary results obtained bring to light links between POS-tags, discourse types and texts annotated by a human expert; these relations are however still incomplete and difficult to interpret from the point of view of linguistics.

Introduction

Depuis de nombreuses années, les méthodes de classification (supervisée ou non) automatique de textes se développent et particulièrement depuis la création du web qui met à notre disposition un nombre incommensurable de textes. Une grande partie des recherches est axée sur l'amélioration des moteurs de recherche qui tentent d'appréhender le contenu des pages web. Ces recherches débouchent également sur des applications de sciences humaines ; elles permettent de découvrir ou de redécouvrir des structures et leur caractérisation, ainsi que d'automatiser les traitements.
Par exemple, depuis les travaux de Biber (1988), de nombreuses recherches se sont développées dans la catégorisation automatique de textes selon leur genre (voir aussi : Karlgren et Cutting 1994, Kessler et al. 1997 et Malrieu et Rastier 2001). Ces travaux utilisent principalement les catégories morphosyntaxiques (CMS) qui peuvent être attribuées automatiquement aux mots à l'aide de logiciels développés à cet effet.

Dans cette même logique, ce projet a pour but de catégoriser automatiquement des

propositions de texte en types de discours pertinents pour la linguistique textuelle, tels que narratifs, explicatifs, argumentatifs, etc. Le corpus d'étude consiste en trois contes de Maupassant. Ceux-ci sont d'abord annotés manuellement par un expert humain. Ensuite, il s'agit d'identifier des structures statistiques pertinentes dans les données, permettant d'effectuer une classification non supervisée. Les étapes types, comme dans tout traitement automatique, sont l'utilisation d'un ou plusieurs algorithmes, suivie de l'interprétation, puis de l'évaluation des résultats. Cette dernière étape peut présenter des difficultés dans le cas des méthodes de classification non supervisée.
Plus spécifiquement, les CMS sont attribuées aux mots du texte. Ensuite, les propositions, telles que segmentées par l'expert humain, sont regroupées, selon les CMS qu'elles contiennent, par un *algorithme* de classification non supervisée. Finalement, les résultats obtenus automatiquement sont *interprétés* et comparés avec ceux produits par l'expert humain pour les *évaluer*.

Il faut encore préciser que cet article n'a pas pour but d'entrer dans les détails des algorithmes et des formules utilisées; le lecteur intéressé par ces points peut se référer à Cocco et al. (2011).

Annotation
Comme il a été expliqué dans l'introduction, la première étape de ce projet a consisté à annoter les textes. Ceci a été fait par un expert humain, étudiant en sciences du langage et de la communication et en français moderne. L'expert a commencé par segmenter le texte en propositions, et c'est cette segmentation manuelle qui sera utilisée dans toute la procédure automatique ; il a en outre annoté les textes en types de discours (voir ci-dessous).

Types de discours
Les types de discours retenus pour ce projet ont été adaptés des travaux de Adam (2008a,b) en linguistique textuelle et de Bronckart (1996) en psycholinguistique et didactique des langues. Les six types finalement choisis ici sont le narratif, l'argumentatif, l'explicatif, le descriptif, le dialogal et l'injonctif. Les cinq premiers sont considérés par Adam (2008a,b) et Bronckart (1996), alors que l'injonctif ne l'est que par Bronckart (1996).

Sans réexposer la théorie relative à ces types de discours, quelques points sont repris ici pour mieux comprendre le but du projet ainsi que les résultats obtenus, et surtout pour expliciter les critères retenus par l'expert humain.
Pour commencer, le type **narratif** correspond au récit raconté. Trois sortes de parties de textes ont été annotées comme étant narratives : *la séquence narrative* (Adam 2008a : 147, schéma 20) composée d'étapes précises, *la période narrative ou l'épisode narratif* où il y a transformation d'un état initial et *le narratif itératif* qui correspond à des actions répétées. Une des principales marques linguistiques du narratif est, généralement, la présence du passé simple, voire de l'imparfait pour le narratif itératif.
Ensuite, pour le type **argumentatif**, *la période et la séquence argumentative* ont été considérées. La première est une suite de propositions liées par des connecteurs argumentatifs et la seconde correspond à des textes ou parties de textes dont le but est de convaincre l'autre de son argument (Adam 2008a : 150, schéma 21). Pour ce projet, les séquences argumentatives incomplètes ont été classifiées comme des périodes argumentatives, et donc annotées comme de type argumentatif. La marque linguistique de ce type est la présence de connecteurs argumentatifs, tels que : *mais, pourtant, cependant, même, d'ailleurs, etc.* (Adam 2008a : 120).
Le type **explicatif**, a pour but d'expliquer quelque chose de non connu, tel le savoir

encyclopédique à transmettre. L'explication répond à la question « Pourquoi ? » (Adam 2008b : 129). A nouveau, ont été annotées comme appartenant à ce type, *les périodes et les séquences explicatives*, les premières étant généralement composées d'une proposition qui pose un problème et d'une explication et les secondes ayant une structure plus complexe (Adam 2008a : 157, schéma 26). La principale marque linguistique est la présence de locutions phraséologiques, telles que : *(si)…c'est parce que/c'est pour que/c'est que, etc.*
Pour le type **descriptif**, il s'agit d'un arrêt sur image où le temps de l'histoire s'arrête et où des propriétés propres à un sujet, qu'il soit animé ou non, lui sont attribuées. Plusieurs marques linguistiques se retrouvent pour ce type : utilisation, en général, de verbes au passé et particulièrement à l'imparfait (à l'exception des cas où la narration ou le discours sont au présent) ; forte proportion d'adjectifs ; présence d'organisateurs spatio-temporels et de verbes d'état et parfois de constructions analogiques.
Bien que le type **dialogal** corresponde à un échange verbal, dans ce projet, le discours direct a aussi été annoté comme étant de type dialogal. En général, ce type de discours se situe à un niveau différent du reste du discours, par exemple dans un système verbo-temporel différent du récit principal. Les marques linguistiques de ce type sont la présence de guillemets et de ponctuation forte, ainsi que le changement de tiroir verbo-temporel. De plus, le discours direct est souvent introduit par des verbes tels que *il dit, elle demanda,* etc.
Finalement, le type **injonctif** est une incitation à l'action : le but de ce type de discours est « de **faire agir** le destinataire d'une certaine manière ou dans une direction donnée » (Bronckart 1996 : 240). Les principales marques de ce type sont l'utilisation de verbes à l'impératif et de verbes introducteurs au dialogue, tels que *il lui ordonna* et, souvent, présence de points d'exclamation. Il faut encore remarquer que dans les textes traités dans ce projet, il se trouve que ce type est toujours inclus dans une séquence ou période dialogale.

Les types décrits ci-dessus ne sont pas univoques, en ce sens que l'interprétation peut, sur certains points, différer d'un expert à l'autre. A ceci s'ajoute le fait que les périodes ou les séquences sont généralement imbriquées les unes dans les autres. En effet, une séquence narrative peut, par exemple, contenir une séquence explicative et/ou descriptive, etc., comme le montre le schéma simplifié de la Fig. 1. Pour pouvoir prendre en compte cette structure hiérarchique, l'expert a utilisé des balises XML pour annoter les textes. Un exemple est donné à la Fig. 2. Mais au final, pour la recherche automatique des types de discours, la structure du texte a été traitée comme étant linéaire, c'est-à-dire que seules les feuilles de l'arbre ont été considérées. Sur la Fig. 1, cela signifie que les propositions 1 à 3 sont considérées comme étant narratives ; les propositions 4 à 5 comme explicatives ; les propositions 6 à 8 comme narratives ; les propositions 9 à 12 comme descriptives ; et la proposition 13 comme dialogale. Avec le même principe, sur la Fig. 2, la proposition de la ligne 6 est considérée comme narrative, les lignes 8 et 9 comme dialogales, les lignes 11 et 12 comme narratives, etc.

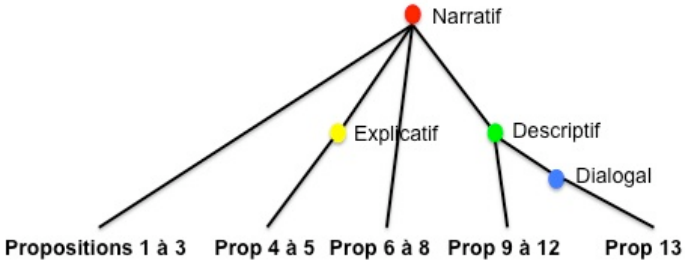


Fig. 1 : Exemple de schéma de la structure d'un texte.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<text source="http://un2sg4.unige.ch/athena/maupassant/maup_fou.html"
date="2011.02.07">
  <title>Un fou ?</title>
<div type="explicatif">
  <div type="narratif">
  <e>Quand on me dit:</e>
    <div type="dialogal">
    <e>"Vous savez</e>
    <e>que Jacques Parent est mort fou dans une maison de santé",</e>
    </div>
  <e>un frisson douloureux, un frisson de peur et d'angoisse me courut le long
des os;</e>
  <e>et je le revis brusquement, ce grand garçon étrange, fou depuis longtemps
peut-être, maniaque inquiétant, effrayant même.</e><cr/>
    <div type="descriptif">
    <e>C'était un homme de quarante ans, haut, maigre, un peu voûté, avec des
yeux d'halluciné, des yeux noirs, si noirs</e>
    <e>qu'on ne distinguait pas la pupille,</e>
    <e>des yeux mobiles, rôdeurs, malades, hantés.</e>
    <e>Quel être singulier, troublant</e>
      <div type="narratif">
      <e>qui apportait, qui jetait un malaise autour de lui, un malaise vague,
de l'âme, du corps, un de ces énervements incompréhensibles</e>
      <e>qui font croire à des influences surnaturelles.</e><cr/>
      </div>
    <e>Il avait un tic gênant: la manie de cacher ses mains.</e>
```

Fig. 2 : Extrait du texte « Un Fou? » annoté par l'expert humain. Les balises <e> et </e> délimitent les propositions.

Corpus

Le corpus qui a été annoté par l'expert humain se compose de trois contes de Maupassant : « L'Orient », « Un Fou ? » et « Un Fou ». Quelques statistiques descriptives concernant ces textes sont données dans le Tab. 1. Il faut préciser que pour le texte « Un Fou », qui est en majorité écrit sous la forme d'un journal intime, les dates ont été retirées, car elles sont difficilement attribuables à l'un des six types de discours.

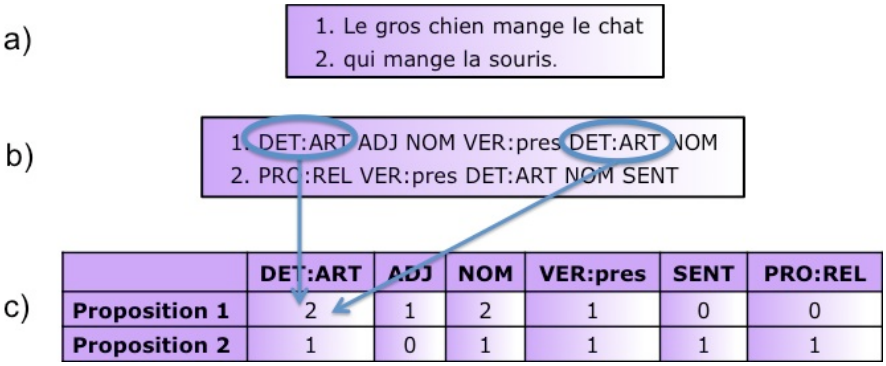
Textes	# phrases	# prop.	# occurences		# formes		% de types de discours selon l'expert humain					
			ponct.	s/ ponct.	mot	CMS	nar	dial	descr	expl	arg	inj
L'Orient	88	189	1'749	1'488	654	27	28.04	25.93	20.11	19.05	4.23	2.65
Un Fou ?	150	314	2'625	2'185	764	28	33.76	14.65	10.51	14.65	18.15	8.28
Un Fou	242	376	3'065	2'548	828	29	42.55	1.86	13.83	11.70	17.82	12.23

Tab. 1 : Statistiques descriptives pour les 3 textes annotés. Nombre de phrases telles que considérées par TreeTagger (Schmid 1994). Nombre de propositions telles que segmentées par l'expert humain. Nombre d'occurrences incluant les ponctuations et les mots composés comme TreeTagger les a tagués. Nombre d'occurrences sans ponctuations ni chiffres et dont les mots composés sont considérés comme des occurrences séparées. Nombre de formes de mots. Nombre de formes de CMS. Les dernières colonnes donnent le pourcentage de propositions pour chaque type de discours (nar = narratif, dial = dialogal, descr = descritptif, expl = explicatif, arg = argumentatif et inj = injonctif).

Partitionnement (classification non supervisée) : méthodologies et algorithmes

Ce qui précédait a été fait manuellement par l'expert humain. Dans cette section nous présenterons les algorithmes effectués par la machine, ainsi qu'un bref rappel sur les différences entre les méthodes de classification supervisée et non supervisée.

Prétraitement des textes
En premier lieu, le texte a dû subir un prétraitement pour pouvoir être ensuite utilisé par les algorithmes de classification non supervisée qui seront présentés ci-dessous. Pour ce faire, on récupère la segmentation faite par l'expert humain (cf. a) Fig. 3). Ensuite, à l'aide de TreeTagger (Schmid 1994), les CMS sont attribuées automatiquement aux mots et ponctuations de chaque proposition (cf. b) Fig. 3). Finalement, une table de contingence est crée (table document-terme). Dans cette table, on compte le nombre de chacune des CMS appartenant à chacune des propositions (cf. c) Fig. 3) et on obtient ainsi la distribution (ou profil) de chaque proposition.



	DET:ART	ADJ	NOM	VER:pres	SENT	PRO:REL
Proposition 1	2	1	2	1	0	0
Proposition 2	1	0	1	1	1	1

Fig. 3 : Prétraitement des textes : a) liste de propositions, b) attribution des CMS par TreeTagger, c) table de contingence des propositions et des CMS.

Méthodes de classification supervisée et non supervisée
Dans les deux cas, il s'agit d'apprentissage automatique, consistant à classer les données textuelles de façon automatique. Dans la classification supervisée, les groupes (ou catégories) dans lesquels on veut classer les données sont connues *a priori*, alors qu'ils ne le sont pas dans le cas non supervisé.
De manière plus précise, pour la classification **supervisée** (« classification » en anglais), on dispose, en entrée, d'un ensemble de données dont on connaît les caractéristiques et les groupes auxquels elles appartiennent. Cette méthode se divise en deux phases. Dans un premier temps (phase d'apprentissage), le programme va apprendre des « règles » de nature statistique sur une partie, conséquente, de l'ensemble des données. Ensuite (phase de test), ce même programme est testé sur les données restantes. Dans cette phase, les données sont transmises au programme sans indication des groupes auxquels elles appartiennent, et il s'agira alors, pour vérifier que le programme fonctionne correctement, de comparer les groupes assignés aux données par le programme aux groupes connus *a priori* pour ces mêmes données.
Dans le cas de la classification **non supervisée** (« clustering » en anglais), les groupes ne sont pas connus *a priori*. Il s'agira donc de partitionner les données de façon à créer des groupes contenant des objets (les propositions ici) similaires (par rapport à leurs caractéristiques qui sont les CMS ici) et tels que les objets appartenant à des groupes différents soient dissimilaires. Ceci correspond à une dissimilarité entre paires d'objets. Concrètement, les propositions sont représentées par des points dans un espace géométrique ; et les groupes, par un ensemble de propositions, chaque groupe étant caractérisé par un centre et une dispersion. Si les dispersions sont égales, les données seront attribuées au centre le plus proche, sinon il faudra aussi tenir compte de la dispersion. Dans le cas le plus simple, les dispersions ne sont pas prises en compte, ce qui a été fait ici. La mesure de distance adoptée est celle du chi-carré, basée sur la différence entre la distribution d'une proposition (cf. c) Fig. 3) et la distribution

du centre.

Dans ce projet, le corpus n'est pas très étendu et il a été choisi d'utiliser des méthodes de classification non supervisée.

Classification non supervisée : K-means et classification floue non supervisée
Deux méthodes ont été utilisées : l'algorithme du K-means et l'algorithme de classification floue non supervisée (fuzzy clustering).

Le principe de l'algorithme du **K-means** est le suivant (Fig. 4) : un nombre *k* de centres sont positionnés au hasard (1), puis les propositions les plus similaires (ou plus proches) de l'un des centres, au sens des distances du chi-carré, sont attribuées au groupe associé à celui-ci (2). Ensuite, de manière itérative, les centres sont repositionnés sur la position moyenne des données attribuées à une certaine classe (3) et les objets sont à nouveau attribués au centre le plus proche (2). L'itération s'arrête lorsque la différence de positions des centres entre les étapes devient négligeable (4). Dans la Fig. 4, une seule itération suffit, ce qui n'est évidemment pas le cas en général.
Il faut encore ajouter qu'ici le K-means est pondéré, c'est-à-dire qu'une proposition plus longue aura plus d'influence sur la position du centre d'une classe.

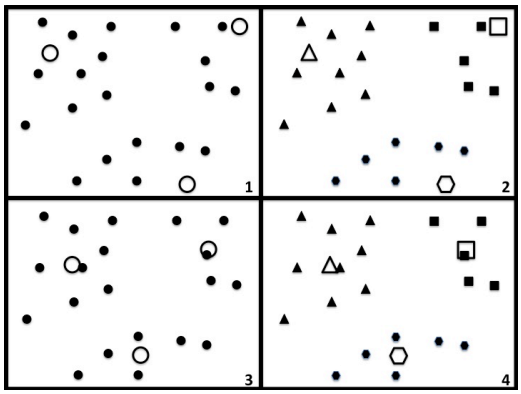


Fig. 4 : Principe du K-means : les documents (les propositions dans ce projet) sont représentés par les formes remplies et les centres des groupes correspondant par les formes creuses.

Pour l'algorithme de **classification floue non supervisée**, le principe est très similaire à celui du K-means dans sa version pondérée, si ce n'est que les propositions (objets) peuvent plus ou moins appartenir, à chaque itération, à chacun des groupes. La Fig. 5 montre le principe de cette appartenance multiple d'un objet à trois groupes qui seraient représentées par les couleurs rouge, vert et bleu. La somme de *b*, *v* et *r* est constante pour chaque point du triangle, et de ce fait, *b* représente le pourcentage de bleu, *v*, le pourcentage de vert et *r*, le pourcentage de rouge qu'il faut mélanger pour obtenir la couleur du point à l'intersection de ces coordonnées. Un exemple plus concret est donné à la Fig. 6 où l'on trouve à nouveau trois groupes représentés par les trois mêmes couleurs. Les points dont les couleurs sont relativement « pures » appartiennent majoritairement au groupe de cette même couleur, alors que les points dont les couleurs sont plus mélangées appartiennent un peu à chacun des trois groupes.
Quel que soit le nombre de groupes en jeu, l'algorithme se termine en attribuant, après stabilisation des appartenances, chaque proposition au groupe dont le centre est le plus proche, sans tenir compte des dispersions.
Pour plus de détails sur cet algorithme, se référer à Cocco et al. (2011), Bavaud (2009) et

Rose et al. (1990).

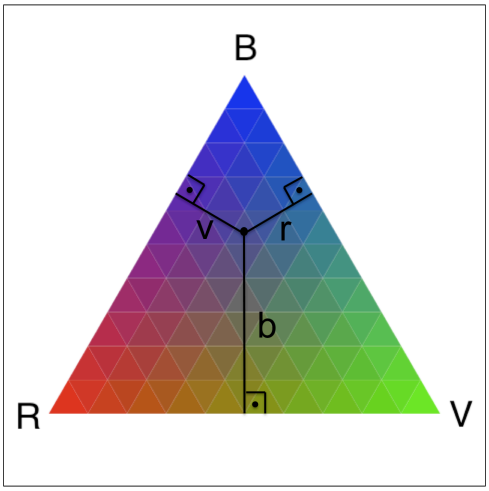


Fig. 5 : Triangle chromatique pour le rouge (R), le vert (V) et le bleu (B).

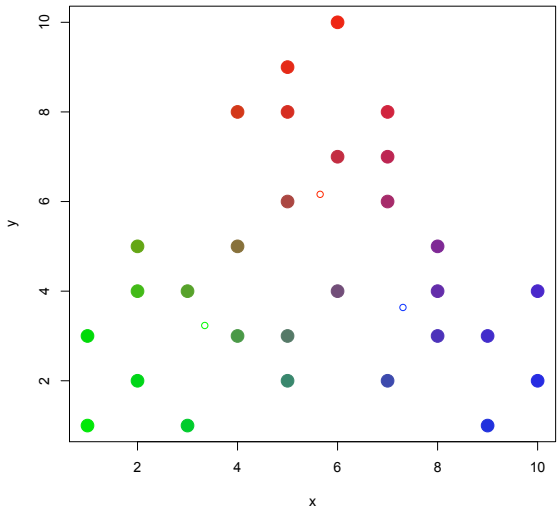


Fig. 6 : Exemple de classification floue non supervisée pour trois groupes (rouge, vert et bleu) : les cercles pleins représentent les objets et les petits cercles vides, les centres.

Il faut encore préciser que comme les centres sont choisis au hasard au début des deux algorithmes présentés ci-dessus, lancer deux exécutions de l'algorithme ne donnera pas deux fois exactement le même résultat, à moins qu'il s'agisse d'un jeu de données très simple ou très caractéristique.

Visualisation des résultats

Il est intéressant, au final, de pouvoir visualiser la table document-terme, donnant les effectifs croisés des propositions et des CMS. En général, les propositions, tout comme les CMS, sont représentées dans un espace de grande dimension, contrairement aux exemples simplifiés qui ont été proposés ci-dessus (Fig. 4 et Fig. 6) qui étaient à deux dimensions. Pour permettre la visualisation par l'œil humain, il faut projeter la table document-terme dans un espace de plus basse dimensionnalité, ce qui peut être obtenu par la technique dite d'analyse factorielle des correspondances (AFC). Un exemple simplifié est donné à la Fig. 7, qui présente une configuration bidimensionnelle (x_1 et x_2) qu'il s'agit de projeter dans une seule dimension en perdant le moins d'information possible. Le nouvel axe optimal en ce sens sera f_1, qui est

caractérisé par le fait qu'il passe au plus près de toutes les données au sens des moindres carrés. Dans le cas général de plus de deux dimensions initiales, on déterminerait ensuite un second axe f_2 qui serait perpendiculaire au premier (flèche en traits-tillés sur la Fig. 7).

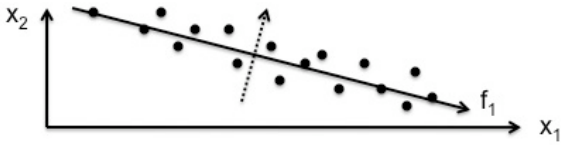


Fig. 7 : Projection d'une configuration bidimensionnelle dans un espace unidimensionnel.

Critères d'évaluation

Une fois obtenus les résultats de la classification, il s'agit d'en évaluer la qualité. A l'inverse des classifications supervisées qu'il est relativement simple d'évaluer, en comparant simplement la partition obtenue automatiquement avec celle attendue (celle de l'expert humain par exemple), l'évaluation des résultats d'une classification non supervisée est plus complexe. En effet, pour cette dernière, on obtient des groupes anonymes (Tab. 3 et Tab. 4), qu'on ne sait pas mettre en comparaison directe avec les groupes proposés par l'expert humain.

Pour pouvoir tout de même évaluer les résultats, on a utilisé deux types de mesures. La première, basée sur le test du chi carré[1], permet de dire si il y a une dépendance (ou non) entre deux variables, qui sont ici la partition automatique obtenue par l'algorithme et celle faite par l'expert humain. D'autre part, on mesurera l'accord de partitions par l'indice de *Jaccard* (*J*) (voir par exemple Denoeud et Guénoche 2006 ; Youness et Saporta 2004) et par l'indice de *Rand corrigé* (*RC*) (voir par exemple Hubert et Arabie 1985 ; Denoeud et Guénoche 2006). Le principe de ces deux mesures est que, pour chaque paire de propositions distinctes, on inspecte si elles sont ou non dans le même groupe dans les deux partitions (celle automatique et celle de l'expert humain). La principale différence entre ces deux indices est que le second soustrait ce qui aurait été obtenu pour une attribution des classes au hasard. Il faut encore préciser que l'indice de Jaccard peut prendre des valeurs comprises entre 0 et 1 (bornes incluses) et que l'indice du Rand corrigé peut également prendre des valeurs négatives.

Partitionnement des groupes : interprétation et évaluation

Dans cette section, les résultats produits pas les algorithmes vont être présentés, interprétés et évalués.

Quotient d'indépendance

Avant de passer aux résultats obtenus avec les algorithmes de classification supervisée, il est intéressant de commencer par observer si, dans les données, il existe bien un lien entre les types de discours et les CMS et si l'on retrouve les principales marques linguistiques correspondant à chaque type. Pour ce faire, il est possible de calculer les quotients d'indépendance (*q*) qui mesurent l'écart à l'indépendance. Plus précisément, le quotient

[1] Pour rappel, la valeur du chi2 mesure la dépendance ou l'association entre les lignes et les colonnes de tableaux croisés du type du Tab. 3, que l'on compare aux degrés de liberté (ddl), mesurant le nombre de catégories présentes. Il en résulte un significativité p, mesurant la probabilité que l'association observée puisse être obtenue par hasard. Par exemple, si $p = 10^{-16}$, cela veut dire qu'il y a une chance sur 10^{16} pour que cette dépendance soit fortuite, donc pratiquement aucune chance. Lorsque la valeur p est faible, on dit que l'association est significative.

d'indépendance est le rapport entre les effectifs observés (un nombre de propositions pour une CMS donnée et un type de discours donné) et les effectifs auxquels on se serait attendu si aucun lien n'existait entre CMS et type de discours. Lorsque le quotient d'indépendance *q* est plus grand que 1, il y attraction mutuelle entre la CMS et le type de discours ; lorsqu'il est plus petit que 1, il y a répulsion mutuelle ; et lorsqu'il est proche de 1, neutralité mutuelle. Les quotients d'indépendance pour les 3 textes étudiés sont donnés dans le Tab. 2.

	« L'Orient »						« Un Fou? »						« Un Fou »					
	nar	dial	descr	expl	arg	inj	nar	dial	descr	expl	arg	inj	nar	dial	descr	expl	arg	inj
ABR													2.59	0.00	0.00	0.00	0.00	0.00
ADJ	0.93	1.15	0.93	1.00	0.34	***1.67***	0.93	0.39	***1.90***	1.08	1.11	0.18	0.68	1.37	***2.03***	1.06	0.77	0.93
ADV	1.56	0.79	0.64	1.09	1.02	1.01	0.77	1.31	0.60	1.43	1.13	0.95	0.94	1.74	0.86	0.86	0.92	1.56
DET:ART	0.84	1.05	1.19	0.95	1.18	0.00	1.02	0.77	1.35	0.45	1.36	0.78	0.90	0.50	0.96	1.11	1.17	1.04
DET:POS	0.55	1.24	0.40	0.86	2.31	5.67	1.70	0.00	0.38	0.69	0.93	0.95	1.09	1.36	1.75	0.90	0.70	0.20
INT	0.00	***3.09***	0.00	0.00	0.00	0.00	0.23	2.06	0.00	1.09	1.22	***5.56***	1.65	***3.72***	0.00	1.53	0.68	0.00
KON	1.11	1.19	0.50	0.92	***1.37***	1.35	0.92	0.83	0.69	***1.51***	1.21	0.17	0.89	0.50	0.99	0.91	***1.22***	1.09
NAM	0.77	0.77	1.25	1.79	0.00	0.00	1.55	1.03	0.56	0.00	0.61	2.78	0.32	0.00	2.27	1.26	0.00	3.33
NOM	0.79	1.03	1.24	0.89	1.24	0.80	1.06	0.56	1.18	0.76	1.23	1.04	0.86	0.64	1.12	0.98	1.16	1.06
NUM	1.43	0.77	1.25	0.34	2.40	0.00	1.38	0.00	1.79	0.00	1.46	0.00	0.71	0.00	1.10	3.68	0.82	0.00
PRO													2.59	0.00	0.00	0.00	0.00	0.00
PRO:DEM	0.63	0.64	1.73	1.67	0.00	0.00	0.20	1.56	1.04	1.51	1.76	0.64	1.25	0.00	0.52	1.40	1.01	0.61
PRO:IND	0.26	1.55	0.84	1.49	0.00	0.00	1.01	0.43	0.94	1.71	1.03	0.00	0.68	0.00	1.32	0.88	1.96	0.00
PRO:PER	1.75	0.83	0.61	0.89	0.94	0.58	1.12	1.21	0.75	1.19	0.75	0.60	1.36	1.12	0.46	0.89	0.94	0.75
PRO:REL	0.51	1.03	1.39	1.49	0.00	0.00	0.74	1.61	0.44	1.27	1.31	0.54	0.77	0.00	1.42	1.11	0.99	1.25
PRP	0.90	1.06	0.86	1.15	1.17	0.86	1.22	0.78	0.93	0.94	0.93	0.50	0.81	0.24	1.54	0.79	1.15	0.87
PRP:det	0.43	1.34	1.51	0.61	1.09	0.00	0.92	0.92	1.33	0.72	1.44	0.00	0.61	0.00	0.85	0.63	1.34	2.36
PUN	0.96	0.96	1.24	0.79	0.98	1.51	1.01	0.88	0.96	0.81	1.11	1.50	0.90	1.96	1.25	0.79	0.87	1.26
PUN:cit	0.00	***2.21***	0.48	0.00	1.83	4.50	0.00	2.06	1.49	0.00	0.00	***12.98***	0.00	***61.30***	0.00	0.00	0.00	0.00
SENT	1.15	0.77	1.03	1.16	0.87	1.61	0.72	1.60	1.13	1.00	0.88	1.93	1.29	2.28	0.55	1.30	0.76	0.70
VER:cond	1.53	0.00	0.00	3.59	0.00	0.00	0.00	2.36	0.00	0.93	2.78	0.00	1.73	4.09	0.00	0.00	1.20	0.00
VER:futu	0.00	2.65	0.24	0.26	0.92	0.00	0.00	4.13	0.00	3.26	0.00	0.00	1.56	0.00	0.61	0.00	0.00	2.67
VER:impf	1.53	0.52	***1.67***	0.90	0.00	0.00	1.38	0.50	***2.85***	0.39	0.30	0.00	1.30	1.14	***1.91***	0.19	0.59	0.16
VER:infi	0.39	0.71	0.72	2.61	1.10	0.00	1.34	1.44	0.78	1.22	0.30	0.00	0.95	0.00	1.03	1.35	1.17	0.63
VER:pper	1.00	1.26	0.78	0.84	1.20	0.00	0.90	1.54	2.08	1.51	0.11	0.00	1.40	0.00	0.71	0.73	0.94	0.56
VER:ppre	0.00	1.77	0.72	1.54	0.00	0.00	2.10	0.39	0.43	0.62	0.23	0.00	0.58	0.00	2.69	0.00	0.50	1.97
VER:pres	1.05	0.58	1.31	1.19	0.95	2.32	0.40	1.97	0.30	1.42	1.23	2.50	1.02	0.28	0.46	1.46	1.00	1.44
VER:simp	***4.59***	0.00	0.00	0.00	0.00	0.00	***2.33***	0.14	0.61	0.00	0.33	0.00	***2.59***	0.00	0.00	0.00	0.00	0.00
VER:subi							0.50	0.00	0.81	4.74	0.00	0.00						
VER:subp	0.00	3.09	0.00	0.00	0.00	0.00	0.00	8.25	0.00	0.00	0.00	0.00	0.52	0.00	0.00	4.05	0.90	1.78

Tab. 2: Quotients d'indépendance pour les 3 textes[2] (les lignes vides signifient que selon TreeTagger, cette CMS ne se trouve pas dans le texte).

Les quotients d'indépendance obtenus montrent qu'il y a attraction mutuelle entre les adjectifs et le type descriptif pour les textes « Un Fou ? » et « Un Fou », et l'injonctif pour le texte « L'Orient ». De même, il y a attraction mutuelle entre les interjections et le dialogal pour « L'Orient » et « Un Fou », et l'injonctif pour « Un Fou ? ». Les conjonctions sont en attraction mutuelle avec l'argumentatif pour les textes « L'Orient » et « Un Fou », et l'explicatif pour « Un Fou ? ». La ponctuation de citation est en attraction mutuelle avec le dialogal pour les textes « L'Orient » et « Un Fou », et avec l'injonctif pour le texte « Un Fou ? ». Finalement, pour les trois textes, il y a attraction mutuelle entre les verbes à l'imparfait et le descriptif, et entre les verbes au passé simple et le narratif.
Il semble donc que les types de discours contiennent effectivement, en général, les CMS

[2] Une explication complète des noms des CMS de ce tableau se trouve sur http://www.ims.uni-stuttgart.de/~schmid/french-tagset.html

auxquelles on s'attendait (marques linguistiques). Il faut tout de même se garder de sur-interpréter ces quotients d'indépendance. Par exemple, la grande valeur obtenue pour la ponctuation de citation et le dialogal pour le texte « Un Fou » est la conséquence de l'apparition exclusive de cette CMS (laquelle n'a d'ailleurs que quatre occurrences dans l'ensemble du texte) pour ce type. De plus, seule la structure linéaire du texte a été exploitée, ce qui pourrait expliquer certaines attractions ou répulsions inattendues.
Pour vérifier qu'il y a réellement une dépendance entre les CMS et les types de discours, les chi carrés ont été calculés sur la table des effectifs observés, selon le même principe qu'expliqué pour les critères d'évaluation, pour les trois textes (« L'Orient » : chi2 = 304.15, ddl = 130, $p = 5.46 \times 10^{-16}$; « Un Fou ? » : chi2 = 672.01, ddl = 135, $p < 10^{-15}$; « Un Fou » : chi2 = 586.63, ddl = 140, $p < 10^{-15}$), ce qui démontre un lien extrêmement fort entre les CMS et les types de discours.

Classification non supervisée
Les exemples qui suivent concerneront uniquement le texte « Un Fou ? », dont les résultats sont les plus intéressants au regard des critères d'évaluation. Sur la Fig. 8, les CMS sont représentés bi-dimensionnellement, avec une inertie totale de 18.5 %. Dans le même espace, la Fig. 9 montre un résultat obtenu avec l'algorithme du K-means pour 6 groupes (dont le tableau croisé est donné dans le Tab. 3) ; et la Fig. 10, un résultat avec l'algorithme de classification floue non supervisée (dont le tableau croisé est donné dans le Tab. 4). Pour les deux algorithmes, il s'agit d'un résultat possible parmi d'autres, car comme il a été expliqué dans la section concernant les méthodes, les centres sont choisis au hasard au départ.

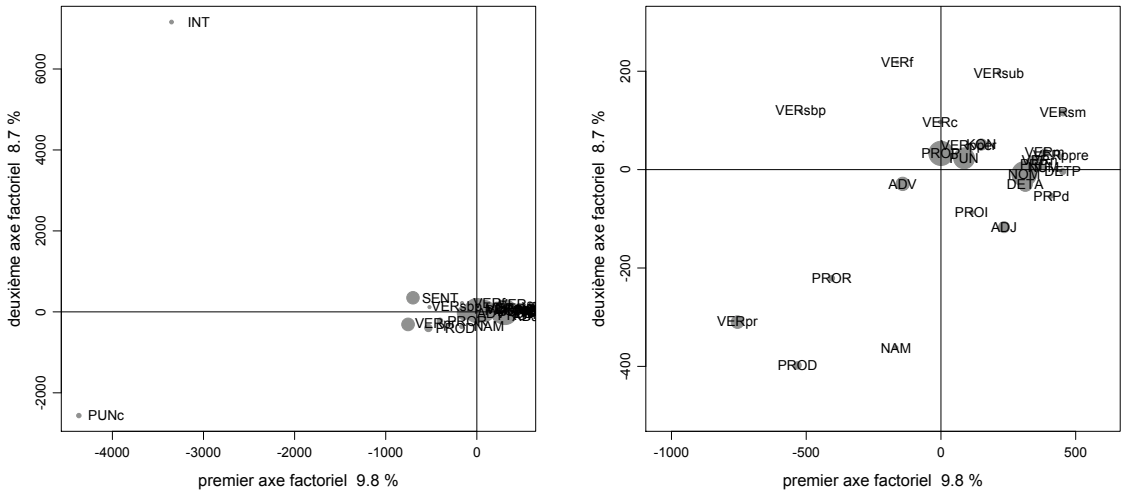


Fig. 8 : Visualisation des CMS pour le texte « Un Fou ? » (à gauche : ensemble des CMS ; à droite : zoom de la figure de gauche).

Comme les CMS (Fig. 8) et les propositions (Fig. 9 et Fig. 10) sont représentées dans le même espace, il est possible de visualiser les relations entre ces dernières. Par exemple, la proposition qui se trouve en A sur la Fig. 9 (coordonnées : -4250, -2100) devrait contenir des points de citations (mêmes coordonnées sur la Fig. 8), et en effet, la proposition est : « « Apporte ! » » qui contient deux points de citations. De même, on s'attend à ce que la proposition au point B sur la Fig. 9 contienne des pronoms démonstratifs (Fig. 8), ce qui bien est le cas : « c'est ceci, ». Finalement, la proposition au point C sur la Fig. 9 contient, comme le montre la Fig. 8, des pronoms démonstratifs et des verbes au présent : « c'est ce qu'on fait partout, ».

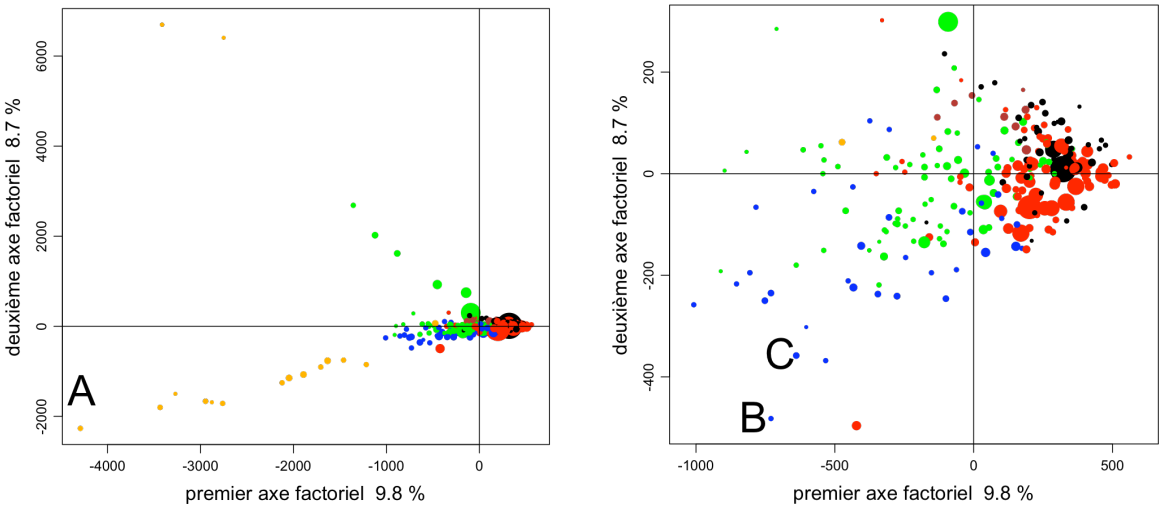


Fig. 9 : Visualisation des propositions pour le texte « Un Fou ? » partitionnées par l'algorithme du K-means en 6 groupes : chaque point représente une proposition dont la taille dépend de la longueur de cette proposition ; et chaque couleur, un groupe (à gauche : ensemble des propositions ; à droite : zoom de la figure de gauche).

		Classes identifiées par l'expert humain						**Total**
		argumentatif	descriptif	dialogal	explicatif	injonctif	narratif	
Classes trouvées auto-matique-ment	1	19	9	24	19	7	16	**94**
	2	24	17	5	7	6	34	**93**
	3	2	2	7	0	12	0	**23**
	4	0	1	0	7	0	2	**10**
	5	10	1	8	10	1	8	**38**
	6	2	3	2	3	0	46	**56**
Total		**57**	**33**	**46**	**46**	**26**	**106**	***314***

Tab. 3 : K-means pour « Un Fou ? » et six groupes : tableau croisé des classes trouvées automatiquement et de celles annotées par l'expert humain.

Le résultat obtenu avec le K-means (Fig. 9 et Tab. 3) démontre une très forte dépendance entre les classes trouvées automatiquement et celles de l'expert humain (chi2 = 201.68, ddl = 25, $p < 10^{-15}$). Les valeurs des indices de similarités de partition sont cependant moins élevés qu'on aurait pu le souhaiter ($J = 0.17$ et $RC = 0.098$).

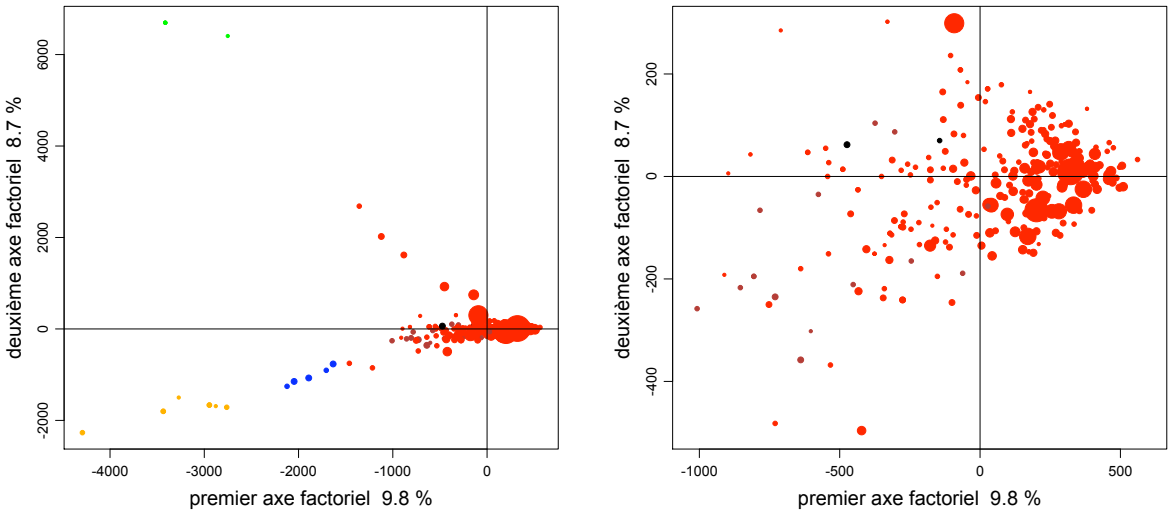


Fig. 10 : Visualisation des propositions pour le texte « Un Fou ? » partitionnées par l'algorithme de classification floue non supervisée en 6 groupes : chaque point représente une proposition dont la taille dépend de la longueur de cette proposition ; et chaque couleur, un groupe (à gauche : ensemble des propositions ; à droite : zoom de la figure de gauche).

		Classes identifiées par l'expert humain						**Total**
		argumentatif	descriptif	dialogal	explicatif	injonctif	narratif	
Classes trouvées auto-matique-ment	1	2	0	1	0	3	0	**6**
	2	52	31	36	41	15	103	**278**
	3	0	2	2	0	4	0	**8**
	4	3	0	4	5	0	3	**15**
	5	0	0	1	0	4	0	**5**
	6	0	0	2	0	0	0	**2**
Total		**57**	**33**	**46**	**46**	**26**	**106**	***314***

Tab. 4 : Classification floue non supervisée pour « Un Fou ? » et six groupes : tableau croisé des classes trouvées automatiquement et de celles annotées par l'expert humain.

Pour la classification floue non supervisée, comme pour le K-means, les critères d'évaluation confirment une très forte dépendance entre les deux partitions (chi2 = 101.05, ddl = 25, $p = 4.18 \times 10^{-11}$), mais, de nouveau les indices de similarité de partitions sont un peu décevants ($J = 0.22$ et $RC = 0.043$).

Au vu des résultats, qui pour rappel seraient différents si l'on exécutait à nouveau les algorithmes de nature probabiliste, il semblerait que le K-means tende à créer des groupes de taille plus homogène que la classification floue non supervisée. Reste le fait que, pour les deux algorithmes, il est difficile de pouvoir mettre en correspondance les classes obtenues automatiquement et celles annotées par l'expert humain.

Conclusions et futures étapes

Ces premiers résultats montrent qu'il existe bien un lien entre les CMS et les types de discours annotés par l'expert humain (quotients d'indépendance), même en considérant la structure du texte comme linéaire plutôt qu'emboîtée. De plus, il existe également une dépendance entre les catégories trouvées automatiquement et celles déterminées par l'expert humain (selon le chi2). Toutefois, cette dernière relation est encore imparfaite (au vu de la

valeur des indices de Jaccard et du Rand corrigé) et son interprétation n'est pas directe. Il faut donc envisager de nouvelles méthodes pour améliorer ces résultats. Par exemple, il serait possible d'utiliser des bi- ou trigrammes, constitués de suites de deux ou trois CMS, plutôt que des unigrammes comme il l'a été fait ici. De plus, il pourrait être intéressant d'ajouter les marques linguistiques typiques de chaque type de discours aux profils de CMS. Il serait aussi possible d'utiliser d'autres mesures de dissimilarité que celle du chi carré, ainsi que de combiner l'algorithme du K-means et celui de la classification floue non supervisée. Une toute autre piste pourrait être d'utiliser la classification supervisée, ce qui permettrait d'obtenir des classes nommées plutôt qu'anonymes. Dans ce cas, il serait souhaitable de disposer d'un plus grand nombre de textes annotés manuellement et il serait alors intéressant de commencer par sélectionner les CMS les plus caractéristiques de chaque type de discours. Enfin, il faudrait pouvoir segmenter le texte en propositions de façon automatique et prendre en compte la structure hiérarchique de ce texte, selon des algorithmes qui resteraient à construire.

Bibliographie

ADAM Jean-Michel (2008a), *La linguistique textuelle : Introduction à l'analyse textuelle des discours*, Paris, Armand Colin, 2ème éd.

ADAM Jean-Michel (2008b), *Les textes : types et prototypes*, Paris, Armand Colin, 2ème éd.

BIBER Douglas (1988), *Variation across Speech and Writting,* Cambridge, Cambridge University Press.

BAVAUD François (2009), « Aggregation Invariance in General Clustering Approaches », *Advances in Data Analysis and Classification*, vol.3, N° 3, pp. 205-225.

COCCO Christelle et al. (2011), « Segmentation and Clustering of Textual Sequences : a Typological Approach », in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, Hissar, Bulgaria, pp. 427-433.

BRONCKART Jean-Paul (1996), *Activité langagière, textes et discours : Pour un interactionisme socio-discursif*, Lausanne ; Paris, Delachaux et Niestlé.

DENOEUD Lucile et GUÉNOCHE Alain (2006), « Comparison of Distance Indices between Partitions », in *Data Science and Classification*, Berlin ; Heidelberg, Springer, pp. 21-28.

DE MAUPASSANT Guy, « L'Orient », *Le Gaulois*, 13 septembre 1883, http://un2sg4.unige.ch/athena/selva/maupassant/textes/orient.html, Thierry Selva, consulté le 5 mars 2011.

DE MAUPASSANT Guy, « Un Fou ? », *Le Figaro*, 1er septembre 1884, http://un2sg4.unige.ch/athena/maupassant/maup_fou.html, Thierry Selva, consulté le 7 février 2011.

DE MAUPASSANT Guy, « Un Fou », *Le Gaulois*, 2 septembre 1885, http://un2sg4.unige.ch/athena/selva/maupassant/textes/unfou.html, Thierry Selva, consulté le 26 avril 2011.

HUBERT Lawrence et ARABIE Phipps (1985), « Comparing Partitions », *Journal of Classification*, vol.2, N° 1, pp. 193-218.

KARLGREN Jussi et CUTTING Douglass (1994), « Recognizing Text Genres with Simple Metrics Using Discriminant Analysis », in *Proceedings of the 15th conference on Computational linguistics,* Kyoto, Japan.

KESSLER Brett et al. (1997), « Automatic Detection of Text Genre », in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* Madrid, pp. 32-38.

MALRIEU Denise et RASTIER François (2001), « Genres et variations

morphosyntaxiques », *Traitement Automatique des Langues*, vol. 42, N° 2, pp. 548-577.

ROSE Kenneth et al. (1990), « Statistical Mechanics and Phase Transitions in Clustering », *Physical Review Letters*, vol. 65, N° 8, pp. 945-948.

SCHMID Helmut (1994), « Probabilistic part-of-speech tagging unsing decision trees », in *Proceedings of the 14th International Conference on Machine Learning,* pp. 44-49.

YOUNESS Genane et SAPORTA Gilbert (2004), « Une méthodologie pour la comparaison de partitions », *Revue de statistique appliquée*, vol. 52, N° 1, pp. 97-120.