Glossaire

Analyse en Composantes Principales (ACP) : *méthode factorielle (Pearson 1901, Hotelling 1933) permettant de fournir un résumé descriptif (sous forme graphique le plus souvent) d'une population décrite par un ensemble de caractères dont les modalités sont des variables numériques continues. Elle ne traite donc que de données quantitatives.

Analyse Factorielle des Correspondances (AFC) : développée par Benzecri (1969), l'analyse des correspondances est une méthode dédiée au traitement de certains tableaux de données particulières, appelés *tableaux de contingence*. Elle « permet d'observer les éventuelles relations existant entre deux variables nominales. Le tableau de contingence (dit aussi de dépendance, ou tableau croisé) est obtenu en ventilant une population selon deux variables nominales. L'ensemble des colonnes du tableau désigne les modalités d'une variable et l'ensemble des lignes correspond à celles de l'autre variable. De ce fait, les lignes et les colonnes, qui désignent deux partitions d'une même population, jouent des rôles symétriques et sont traités de manière analogue. » (Lebart, 2004)

Apprentissage supervisé : l'apprentissage supervisé consiste à inférer un modèle de prédiction à partir d'un ensemble d'apprentissage, c'est-à-dire plusieurs couples de la forme {observation, étiquette}, où chaque étiquette dépend de l'observation à laquelle elle est associée. Un algorithme d'apprentissage supervisé a pour but de déterminer une fonction s'approchant au mieux de la relation liant les observations et les étiquettes à partir de l'ensemble d'apprentissage uniquement. Cette fonction doit par ailleurs posséder de bonnes propriétés de généralisation et ainsi être capable d'associer une étiquette adéquate à une observation qui n'est pas dans l'ensemble d'apprentissage.

Apprentissage non supervisé : l'apprentissage non supervisé consiste à inférer des connaissances sur des classes sur la seule base des échantillons d'apprentissage, et sans savoir *a priori* à quelles classes ils appartiennent. Contrairement à l'apprentissage supervisé, on ne dispose que d'une base d'entrées et c'est le système qui doit déterminer ses sorties en fonction des similarités détectées entres les différentes entrées (règle d'auto organisation).

Arbres de décision : souvent employés en fouille de données, les arbres de décision permettent de répartir une population d'individus (textes ici) en groupes homogènes, selon un ensemble de variables discriminantes (morphosyntaxiques et lexicales dans nos analyses) et en vue d'un objectif déterminé. Ils produisent une représentation graphique d'une procédure de classification : la présence et la position d'un attribut dans l'arbre indiquent son importance dans le processus de classification ainsi que la classe favorisée par ce dernier. En ce sens, les arbres de décision présentent l'intérêt d'être intuitifs et

facilement interprétables. Dans nos expérimentations, c'est l'algorithme de classification supervisée C4.5 que nous avons utilisé.

Boostrap (rééchantillonnage) : l'épreuve du bootstrap (Efron, 1979 et Efron et Diaconis 1981) est un test extrêmement sévère pour évaluer la stabilité d'une structure. Elle consiste à faire de l'inférence statistique sur de « nouveaux » échantillons tirés à partir d'un échantillon initial. Dans le cadre de notre travail, on a utilisé la méthode du bootstrap pour « *perturber* le corpus de 224 articles de la façon suivante : on place les 224 titres dans une urne, et l'on tire, *avec remise*, 224 fois un titre au hasard. Dans la série de 224 éléments obtenus, certains titres n'apparaissent pas, d'autres apparaissent deux fois ou plus (en moyenne, environ 70% des titres sont présents à l'issue du tirage). Le corpus ainsi perturbé est une *réplication* du corpus initial. D'autres réplications sont construites (une trentaine suffit dans ce contexte), et l'analyse de l'ensemble des réplications définit des halos autour de chaque point, halos qui sont matérialisés par des ellipses de confiance. » (Lebart, rapport de thèse)

Cartes de Kohonen (p. 156) : C'est dans le début des années 80 que T. Kohonen a proposé la méthode des cartes auto-organisatrices (SOM, *Self-Organizing Maps*), qu'on présente généralement comme un cas particulier des réseaux de neurones. L'algorithme de Kohonen est un algorithme de classification qui regroupe les observations en classes, en respectant la topologie de l'espace des observations. Une notion *a priori* de voisinage entre classes est ainsi définie.

On suppose généralement que les classes sont disposées sur une grille rectangulaire aux mailles déformables dans laquelle les voisins de chaque classe sont naturellement définis ; en d'autres termes, on construit une représentation bidimensionnelle d'une distribution multidimensionnelle et on dispose d'une représentation graphique unique des données dans l'espace de sortie, ce qui présente un intérêt indiscutable lorsqu'il s'agit d'observer un nombre important de données

Dans la mesure où elles sont capables de s'étirer et d'épouser plus étroitement le nuage de points, les cartes de Kohonen peuvent être considérées comme un équivalent qualitatif et non linéaire de l'ACP.

Classification Ascendante Hiérarchique (CAH) : La CAH est une *méthode de classification qui consiste à regrouper les individus ayant un comportement similaire en classes en fonction de deux critères : les individus d'une même classe sont le plus semblable possible, et les classes sont les plus disjointes possible.

Les méthodes ascendantes sont agglomératives : à chaque étape du processus, on créée une partition en agrégeant deux à deux les individus, ou les groupes d'individus les plus proches. Pour un niveau de précision donné, deux individus peuvent être confondus dans le même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents ; les classes formées étant emboîtantes, la classification est qualifiée de hiérarchique.

Coefficient de variation (p. 102) : écart-type divisé par la moyenne, souvent exprimé comme un pourcentage à la moyenne.

Concept de fond / de forme : distinction élaborée par Rastier (2005) entre les unités participant à l'évolution des *formes* sémantiques le long du texte, et celles appartenant au *fond* sémantique général. Les premières sont caractérisantes, car elles correspondent aux « concepts » effectivement élaborés, débattus et utilisés, alors que les secondes restent d'une grande généralité.

Configuration optative : composante facultative mais caractéristique d'un genre ou d'un discours (*e.g.* l'exemple, la citation, la note de bas de page, etc.).

Configuration tactique : déroulement textuel d'un élément (étiquette morphosyntaxique ou mot), mesuré ici en *déciles de rang* d'occurrences d'éléments par texte (CR, Loiseau), et représenté sous la forme d'un diagramme ordonné.

Champ générique : groupe de genres qui contrastent, voire rivalisent dans un champ pratique. Par exemple, au sein du discours scientifique, les genres de la revue (article, discussion, réponse, compte rendu, présentation de revue, etc.) constituent un champ générique propre.

Discours : ensemble d'usages linguistiques codifiés attaché à un type de pratique sociale. Ex. : discours juridique, scientifique, littéraire.

Domaine : le domaine (ou la discipline) s'inscrit dans une pratique sociale. Il est commun aux divers genres propres au discours qui correspond à cette pratique. Dans un domaine déterminé, il n'existe généralement pas de polysémie.

Genre : niveau normatif de contraintes et de prescriptions positives ou négatives régulant la production et l'interprétation d'un texte. Puisque tout texte relève d'un genre, et que tout genre relève d'un discours, le genre permet de relier les textes aux discours.

Information mutuelle : degré de dépendance d'un couple (X,Y) de variables. L'information mutuelle mesure la quantité d'information apportée en moyenne par une réalisation de X sur les probabilités de réalisation de Y. Dans le cadre de notre étude, nous avons recouru à l'information mutuelle (p. 314) pour ordonner les substantifs de chaque corpus d'observation selon sa probabilité d'apparition dans les textes.

Intervalle de confiance d'Anderson : « Anderson (1963) a calculé les lois limites des valeurs propres d'une *ACP sans nécessairement supposer que les valeurs théoriques correspondantes sont distinctes. L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage supposé laplacien. L'empiétement des intervalles de deux valeurs propres consécutives suggérera donc l'égalité de ces valeurs propres. Les axes correspondants sont définis à une rotation près. Ainsi l'utilisateur pourra éviter d'interpréter un axe instable selon ce critère. » (Lebart *et al.*, 2003 : 205).

Méthodes factorielles : les méthodes factorielles visent à résumer de manière synthétique de vastes ensembles de valeurs numériques par le biais d'un nombre plus restreint de variables artificielles nouvelles, les facteurs. Chaque facteur représente un groupement de traits intercorrélés (linguistiques ici). Les facteurs sont généralement représentés sous forme de visualisations graphiques où les objets à décrire deviennent des points sur un axe ou dans un plan. Contrairement aux méthodes de classification qui font appel à une démarche algorithmique, les méthodes factorielles utilisent des calculs d'ajustement faisant essentiellement appel à l'algèbre linéaire.

Méthodes de classification : les méthodes de classification sont destinées à produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères. Elles mettent en jeu une formulation et des calculs algorithmiques, et produisent des classes ou des familles de classes, permettant de grouper et de ranger les objets à décrire.

L'utilisation conjointe de l'analyse factorielle et de la classification permettra de se prononcer non seulement sur la réalité des classes, mais également sur leurs positions relatives. Le plus souvent, lors des approches exploratoires, les partitions ou les arbres de classification viendront compléter et nuancer des analyses factorielles préalables.

Peeling : épluchage progressif des points aberrants.

Pratique sociale : activité codifiée qui met en jeu des rapports spécifiques entre le niveau sémiotique (dont relèvent les textes), le niveau des représentations mentales et le niveau physique.

Style d'auteur (style personnel) : usage d'un sociolecte propre à un énonciateur. Usage singulier du genre.

SVM (Support Vector Machines/Séparateurs à Vastes Marges) : méthode de classification binaire par apprentissage supervisé. De manière simplifiée, cette méthode consiste à apprendre un classifieur dans un nouvel espace d'attributs de dimension plus importante que l'espace initial. Ce nouvel espace peut-être obtenu par différents types de fonctions noyaux (*e.g.* linéaire, polynomial, RBF, etc. *v.* Vapnik, 1995 pour plus de précisions sur la technique d'apprentissage par SVM). Plusieurs études empiriques (*e.g.* Dumais, 1998) ayant montré que les meilleures performances en classification textuelle sont obtenues avec des SVMs linéaires, c'est ce type de noyau que nous avons retenu dans nos expérimentations. (p. 313)

Texte : suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque. Les textes sont l'objet empirique de la linguistique. (Rastier, 1996)

Tri Systématique de Signification (TSS) : le TSS est une méthode de validation externe permettant de confronter le corpus avec des informations externes concernant des

composantes du corpus en positionnant des variables illustratives *a posteriori* sur les axes principaux au moyen d'une *valeur-test* exprimant la signification statistique de la coordonnée de la variable sur l'axe. (Lebart, 2004, p. 712).