Glossaire pour la statistique textuelle

NB : Les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :

ac	Analyse factorielle des correspondances
acm	Analyse des correspondances multiples
cla	Classification
sp	Méthode des Spécificités
sr	Analyse des segments répétés
ling	Linguistique
stat	Statistique
sa	Segmentation automatique

accroissement spécifique - (sp) spécificité* calculée pour une partie d'un corpus par rapport à une partie antérieure

algorithme - ensemble des règles opératoires propres à un calcul.

analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

analyse des correspondances (stat)- méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou χ^2).

analyse des correspondances multiples (stat) **-** méthode d'analyse des correspondances s'appliquant à l'étude de tableaux disjonctifs complets. Tableaux binaires dont les lignes sont des individus ou observations et les colonnes la juxtaposition des modalités de réponse à des questions (les modalités de réponse à une question s'excluant mutuellement).

caractère (sa) - signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / non-délimiteurs (sa) - distinction opérée sur l'ensemble des caractères, qui entrent dans la composition du texte permettant aux

procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquence** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

classification (stat) - technique statistique permettant de regrouper des individus ou observations entre lesquels a été définie une distance.

classification hiérarchique (cla) - technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

concordance (sa) - l'ensemble de lignes de contexte se rapportant à une même forme-pôle.

contribution absolue (ou contribution) - (ac) contribution apportée par un élément au facteur . Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.

contribution relative (ou cosinus carré) - (ac) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.

cooccurrence (sa) - (une c.) - présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.

corpus (ling) - ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

(lexicométrie) ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.

délimiteurs de séquence - (sa) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général - le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

dendrogramme - (cla) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

discours/langue -

distance du chi-2

éditions de contextes

éléments d'un segment

éléments actifs-

éléments supplémentaires (ou illustratifs)-

énoncé/énonciation

expansion contrainte -

expansion d'un segment

expansion récurrente d'un terme

facteur

forme- forme graphique

forme banale

forme caractéristique

forme commune - forme attestée dans chacune des parties du corpus.

forme originale- (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.

fréquence (sa) - (d'une unité textuelle) le nombre de ses occurrences dans le corpus.

fréquence d'un segment (sr) - (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale (sa) - fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de").

fréquence relative (sa) - la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

gamme des fréquences (sa) **-** suite notée V_k, des effectifs correspondant aux formes de fréquence k, lorsque k varie de 1 à la fréquence maximale.

hapax - gr. hapax (legomenon), "chose dite une seule fois".

(sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

identification - (stat, ling, sa) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

index - (sa) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regouper les références* relatives à l'ensemble des occurrences d'une même forme.

index alphabétique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).

index hiérarchique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.

index par parties - ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.

item de réponse - (ou modalité de réponse) élément de réponse préétabli dans une question fermée.

lemmatisation - regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

lexical - (ling) qui concerne le lexique* ou le vocabulaire*.

lexicométrie ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire* d'un corpus de textes.

lexique - (ling) ensemble virtuel des mots d'une langue.

longueur (sa) - (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. : partie, fragment, etc.). Synonyme de taille.

On note: T la longueur du corpus; t_j celle de la partie (ou tranche) numéro j du corpus.

longueur d'un segment (sr) - le nombre des occurrences entrant dans la composition de ce segment.

noyaux factuels - (cla) classes d'une partition* d'un ensemble d'observations synthétisant une batterie de descripteurs objectifs (sexe, âge, profession, etc.).

occurrence (sa) **-** suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.

ordre lexicographique -

– pour les formes graphiques :

l'ordre selon lequel les formes sont classées dans un dictionnaire.

NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple, rangées dans cet ordre, les formes : *mais, maïs, maison, maître* .

– pour les polyformes:

ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante, les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.

ordre lexicométrique (sa) -

– pour les formes graphiques :

ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes; les formes de même fréquence sont classées par ordre lexicographique.

– pour les polyformes:

ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

paradigme- (ling) ensemble des termes qui peuvent figurer en un point de la chaîne parlée.

paradigmatique- (sa) qui concerne le regroupement en série des unités textuelles, indépendamment de leur ordre de succession dans la chaîne écrite.

partie - (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition - (d'un corpus de textes) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

(d'un ensemble, d'un échantillon) division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.

partition longitudinale - (sa) partition d'un corpus en fonction d'une variable qui définit un ordre sur l'ensemble des parties

périodisation (sa) - regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.

phrase - (sa) fragment de texte compris entre deux séparateurs* de phrase.

places - (sa) pour un texte comptant T occurrences, suite de T objets correspondant chacun à une des occurrences du texte, éventuellement séparés par des délimiteurs* de séquence correspondant aux ponctuations du texte de départ.

polyforme (sr) - archétype des occurrences d'un segment; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

ponctuation - Système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.

(*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.

post-codage - opération manuelle qui consiste à repérer les principales catégories de réponses libres sur un sous-échantillon de réponses, puis à fermer la question ouverte correspondante, en affectant toutes les réponses à ces catégories.

pourcentages d'inertie - (ac ou acm) quantités proportionnelles aux valeurs propres* dont la somme est égale à 100. Notées τ_α.

profil - (stat et ac) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

question fermée - question dont les seules réponses possibles sont proposées explicitement à la personne interrogée.

question ouverte - question posée sans grille de réponse préétablie, dont la réponse peut être numérique (ex: *Quelles doivent être, selon vous, les ressources minimum d'une famille ayant trois enfants de moins de 16 ans?*), ou textuelle (ex: *pouvez-vous justifier votre choix?*).

références (sa) - système de coordonnées numériques permettant de repérer dans le texte d'origine chacune des occurrences issues de la segmentation (ex : le tome, la page, la ligne, la position de l'occurrence dans la ligne) ou de situer rapidement cette occurrence parmi des catégories prédéfinies (auteur, année de parution, citation, mise en valeur, etc.).

répartition (sa) - (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.

réponse modale - (d'une classe d'individus, d'une partie* de corpus*) réponse sélectionnée en fonction de son caractère représentatif d'une classe ou d'une partie en général à partir des formes* caractéristiques qu'elle contient.

segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.

segment répété (sr) - (ou polyforme répétée) suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentaire - (sr) ensemble des termes* attestés dans le corpus.

segmentation - opération qui consiste à délimiter des unités minimales* dans un texte.

segmentation automatique - ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.

séparateurs de phrases - (sa) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur* de séquence.

seuil - (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

seuil d'absence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence A(j) pour laquelle toute forme de fréquence supérieure à

A(j) dans le corpus et absente dans la partie j est spécifique (négative) pour la partie j.

seuil de présence spécifique - (sp) pour un seuil fixé, pour une partie du corpus fixée, fréquence B(j) pour laquelle toute forme de fréquence inférieure à A(j) dans le corpus et présente dans la partie j est spécifique (positive) pour la partie j.

sous-fréquence (sa) - (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.

sous-segments (sr) - pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.

spécificité chronologique - (sp) spécificité* portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale*.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

stock distributionnel du vocabulaire - (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.

syntagmatique- (sa) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

syntagme- (ling) groupe de mots en séquence formant une unité à l'intérieur de la phrase.

tableau de contingence (stat) - synonyme de tableau de fréquences ou de tableau croisé: tableau dont les lignes et les colonnes représentent respectivement les modalités de deux questions (ou deux variables nominales) , et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.

tableau lexical entier (TLE) - tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties

du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.

tableau des segments répétés (TSR) - tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

tableau lexical- tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui correspondent à des formes dont la fréquence est inférieure à un seuil donné).

taille- (sa) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme - (sr) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2,3, etc.

termes contraints / termes libres - Un terme S1 est contraint dans un autre terme S2 de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment S2. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

unités minimales (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent).

valeur modale - (stat) valeur pour laquelle une distribution atteint son maximum.

valeurs propres - (ac ou acm) quantités permettant de juger de l'importance des facteurs successifs de la décomposition factorielle. La valeur propre notée λ_α. mesure la dispersion des éléments sur l' axe.α.

valeurs-tests - (ac ou acm) quantités permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur test dépasse 2 en valeur absolue, il y a 95 chances sur 100 que la position de l'élément correspondant ne puisse être due au hasard.

variables actives - variables utilisées pour dresser une typologie, soit par analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids des variables actives, qui doivent de ce fait constituer un ensemble homogène.

variables supplémentaires (ou illustratives) - variables utilisées a posteriori pour illustrer des plans factoriels ou des classes. Une variable supplémentaire peut-être considérée comme une variable active munie d'un poids nul.

variables de type T - var able dont la fréquence est à peu près proport onnelle à l'allongement du texte. (ex : la fréquence max male)

variables de type V- var able dont l'accro ssement a tendance à d m nuer avec l'allongement du texte (ex : le nombre des formes, le nombre des hapax).

ventilation (sa) - (des occurrences d'une un té dans les part es du corpus) La su te des n nombres (n = nombre de part es du corpus) const tuée par la success on des sous-fréquences* de cette un té dans chacune des part es, pr ses dans l'ordre des part es.

vocabulaire (sa) - ensemble des formes* attestées dans un corpus de textes.

vocabulaire commun - (sa) l'ensemble des formes attestées dans chacune des part es du corpus.

vocabulaire de base - (sp) ensemble des formes du corpus ne présentant, pour un seu l f xé, aucune spéc f c té (négat ve ou pos t ve) dans aucune des part es , (.e. l'ensemble des formes qu sont "banales" pour chacune des part es du corpus).

vocabulaire original- (sa) (pour une part e du corpus) l'ensemble des formes* or g nales* pour cette part e.

voisinage d'une occurrence - (sa) pour une occurrence donnée du texte, tout segment (su te d'occurrences consécut ves, non séparées par un dél m teur de séquence) contenant cette occurrence.

voisinages d'une forme - (sa) ensemble const tué par les vo s nages de chacune des occurrences correspondant à la forme donnée.