Concordanciers : Thème et variations

Bénédicte Pincemin, Fabrice Issac, Marc Chanove, Michel Mathieu-Colas

Laboratoire de Linguistique Informatique, FRE2882 CNRS – Université Paris 13 – Avenue Jean-Baptiste Clément – F-93430 Villetaneuse – France

Abstract

The computation of a concordance is usually determined by three parameters : the word (or linguistic pattern) to be found, the size of the context given for each token, and the way the extracts should be sorted in order to facilitate the analysis. The power of this technic lies in the visual effects it creates by aligning and grouping the contexts through the centered-column presentation and the sorting of the lines. These principles can be generalized and extended : the pattern to be found can be decomposed into several zones, and each of them can be aligned as a column, or/and can be sorted.
We illustrate these proposals by the implementation of a concordancer (KWAC-LLI) specialized for corpus linguistics in a distributional semantics approach. The corpus can be analysed according to four strategies, based on the syntagmatic or paradigmatic relation between predicates and arguments. The specialized concordancer tests two new features. The first one is a table which gives a global view of the concordance, with hypertext access to the detailed contexts. The second new feature is a linguistic sort, directly derived from the "classe d'objets" theory.

Résumé

Le calcul d'un concordancier se définit classiquement par trois paramètres : l'expression d'un pivot, la délimitation du contexte donné pour chaque occurrence relevée du pivot, et l'organisation des extraits par un tri facilitant le dépouillement. L'efficacité propre à cette technique tient essentiellement aux effets d'alignement et de regroupement issus de la présentation du pivot sur une colonne et des tris sur le pivot et son environnement. Nous proposons donc une généralisation de la technique des concordances avec l'articulation interne du pivot en plusieurs zones, focalisant et démultipliant les possibilités d'alignement et de tri.
Nous prenons appui sur cette réflexion pour développer un concordancier (KWAC-LLI) adapté aux besoins linguistiques d'une sémantique distributionnelle, en l'occurrence la théorie des classes d'objets. Une combinatoire de quatre stratégies d'exploration de corpus peut être ainsi outillée, selon que l'on part de prédicats ou d'arguments pour rechercher d'autres prédicats ou d'autres arguments. Le concordancier s'enrichit dans ce contexte de deux innovations significatives : la présentation globale et synthétique des résultats sous forme de tableau hypertexte, et le tri des lignes du tableau traduisant directement un critère de pertinence linguistique donné par la théorie des classes d'objets.

Mots-clés : concordances, logiciels d'analyse de données textuelles, linguistique de corpus, sémantique distributionnelle, théorie des classes d'objets.

1. Introduction

Le recours maintenant possible aux corpus numériques ouvre de nouveaux modes d'investigation pour le travail du linguiste. Dans notre laboratoire, nous faisons une description sémantique de la langue pour la réalisation de lexiques électroniques, adaptés aux besoins des traitements automatiques. Cette description a été mise au point par Gaston Gross et son équipe, et elle est souvent désignée par l'un de ses concepts centraux : les *classes d'objets* (Le Pesant et Mathieu-Colas, 1998). Après une étude générale des caractéristiques des concordanciers et une proposition de généralisation qui en reprend et développe les aspects les plus intéressants (liés aux effets visuels de superposition des contextes), notre

propos ici est de présenter une application d'analyse de corpus textuels, qui est une forme renouvelée et originale de concordancier. La technique des concordances y est revue et adaptée non seulement pour pallier certaines difficultés bien connues des praticiens (comme le volume des relevés obtenus), mais aussi pour intégrer à l'application des principes linguistiques fondamentaux de la théorie des classes d'objets, optimisant le repérage des items pertinents pour compléter ou caractériser une classe. Ce travail intéresse bien entendu les linguistes utilisant la théorie des classes d'objets ou d'autres formes d'analyses distributionnelles. Il peut également inspirer les concepteurs d'outils de statistique textuelle, qui y verront une manière de développer les fonctions de concordance en mettant en valeur leurs atouts spécifiques, et de les rendre ainsi plus proches et plus efficaces pour des besoins variés.

2. Concordances informatisées

2.1. Définition par synthèse de l'état de l'art

L'observation des logiciels d'analyse textuelle et les manuels du domaine (comme (Lebart et Salem, 1994)) permettent de définir le calcul d'une concordance par trois paramètres : pivot, taille de contexte, et tri. Soit donc une possible définition comme suit :

> Un corpus étant fixé, une concordance est la liste de toutes les occurrences d'un pivot, alignées verticalement en colonne (nous dirons "empilées"), entourées de part et d'autre par leur contexte, et triées selon un critère pertinent pour l'analyse.

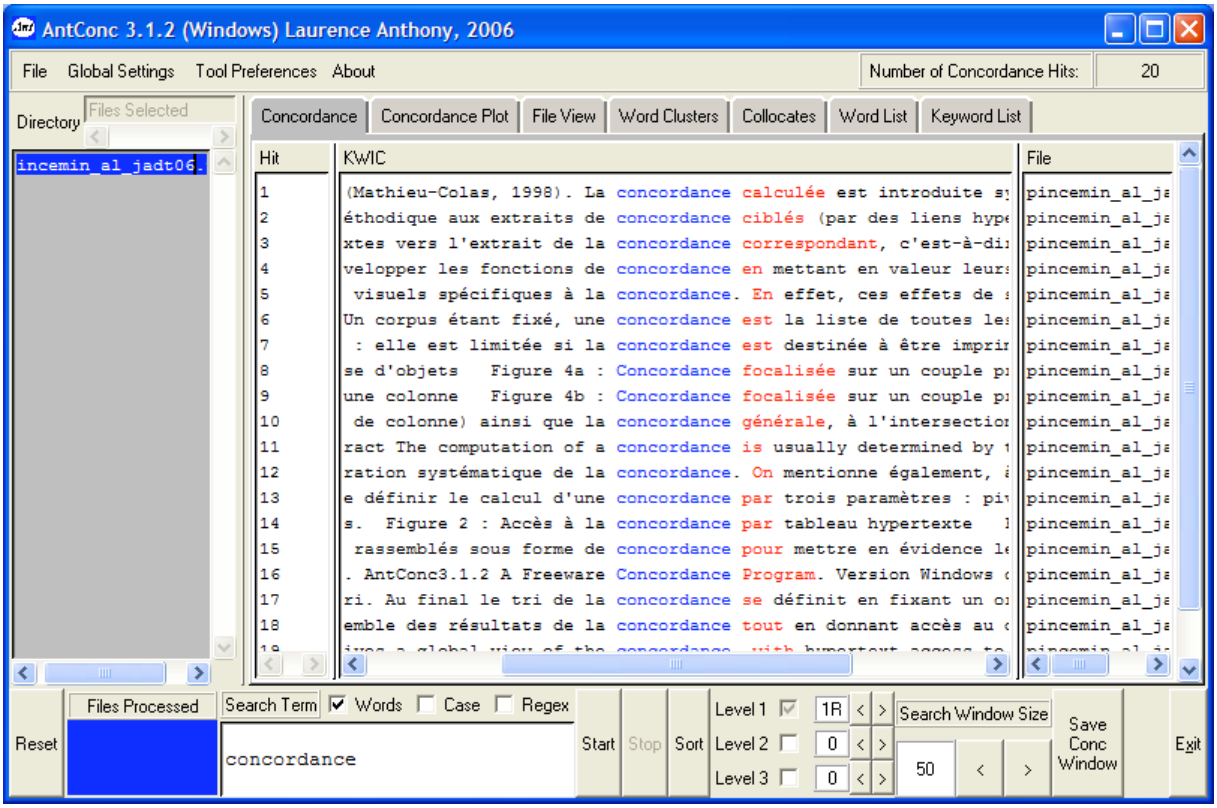


Figure 1 : Exemple de concordancier - ici le concordancier AntConc *(Anthony, 2006)*

2.1.1. Le pivot

Le codage du corpus, la modélisation interne des données dans le logiciel, et la conception de l'interface, déterminent les motifs repérables dans le corpus et mobilisables en tant que pivot. Le pivot typique est un mot, tel que manifesté en corpus par sa graphie. La recherche peut être assouplie en portant sur une chaîne de caractères, ce qui est par exemple une heuristique simple pour neutraliser la terminaison, et est souvent efficace pour viser un même mot par delà ses variations de flexion, ou encore une thématique exprimée par les mots d'une même famille. Dans le même ordre d'idées, le pivot peut être étendu à une liste de mots, afin d'avoir un recensement plus ouvert et plus ciblé des variantes d'expression du concept recherché.

L'affinement du paramétrage du pivot consiste ensuite à lui permettre de recouvrir des structures composées (par exemple des expressions), ou à donner accès à des niveaux de description de nature variée (tels que le lemme ou la catégorie grammaticale, si l'étiquetage du corpus l'a fourni). Ces axes de développement sont combinables, pour permettre par exemple le repérage de constructions linguistiques étendues et précises.

Techniquement, cela peut se traduire, comme dans le langage d'interrogation CQP (Christ, 1994) utilisé par le concordancier de *Weblex* (Heiden, 2002), par une modélisation du corpus en strates d'informations de nature diversifiée, et une puissante syntaxe d'interrogation croisant trois paliers de description (*valeurs* exprimées par des chaînes de caractères et analysables comme telles, *attributs* paradigmatiques déclinant les différentes strates, *positionnement* syntagmatique réglant l'enchaînement en séquence composée) avec les principaux opérateurs (alternative "ou", conjonction "et", joker, exclusion).

2.1.2. Taille du contexte

La taille du contexte correspond usuellement à la longueur de la ligne.

Le logiciel permet parfois de régler la taille de contexte, les besoins variant suivant les recherches : un contexte court (de l'ordre du syntagme, de quelques mots) peut suffire pour une étude à dominante lexicale, un contexte pas trop restreint (de l'ordre de la proposition ou de la phrase) est nécessaire pour des observations d'ordre syntaxique, et un contexte plus étendu (de l'ordre de quelques phrases, du paragraphe) peut être requis pour un point de vue sémantique[1]. Le réglage de la taille du contexte concourt ainsi à mieux focaliser le regard sur les zones les plus pertinentes pour l'étude, à optimiser la sélection de passages pour une haute densité de pertinence.

La longueur de ligne peut également être modulable : elle est limitée si la concordance est destinée à être imprimée, mais théoriquement non bornée pour une visualisation à l'écran, dans une fenêtre munie d'un curseur horizontal. Si le concordancier permet d'ajuster la taille du contexte et celle de la ligne, un cas particulier à prévoir est celui où le contexte est plus long que la ligne : il est difficile alors de préserver l'effet d'empilement, non seulement au niveau du pivot lui-même mais aussi sur son contexte immédiat.

La colonne d'empilement du pivot est généralement centrée au milieu de la page. Dans certains concordanciers, il possible de régler la part accordée à l'un et l'autre des deux contextes (gauche ou droit), généralement en indiquant le nombre de caractères de part et d'autre du pivot.

[1] Cette présentation est très simplifiée ; notamment, lexique, syntaxe et sémantique ne sont pas indépendants.

2.1.3. Tris

Bien que certains outils se dénomment concordanciers sans présenter de possibilité de tris, ce paramètre nous paraît essentiel, car il est une caractéristique originale et puissante de cette technique d'analyse textuelle. En effet, la force de la concordance électronique, c'est bien cette organisation très visuelle des contextes, qui joue sur les effets d'alignement et de répétition pour mettre efficacement en valeur les régularités plus ou moins massives et les écarts significatifs.

Les concordanciers évolués proposent généralement trois ou quatre possibilités de tri fixant l'ordre de présentation des lignes[2] : ordre de déroulement du corpus (dit ordre chronologique), tri alphabétique sur le pivot (dans la mesure où le pivot revêt effectivement diverses formes au fil des occurrences), tri alphabétique sur le contexte droit (qui suit le pivot), tri alphabétique sur le contexte gauche. Ce dernier consiste à trier les contextes selon l'ordre alphabétique du mot qui précède le pivot, puis du mot encore avant, et ainsi de suite (ce n'est donc pas un simple miroir du tri droit, qui opérerait sur la chaîne de caractères globale du contexte gauche et trierait alphabétiquement caractère par caractère en progressant de la droite vers la gauche : elle s'appuie sur un découpage en mots).

Ces diverses possibilités de tri peuvent s'emboîter, et donner lieu à un tri multiple : par exemple, tri sur le pivot, puis (à forme de pivot identique) tri chronologique. Une limite pratique à l'intérêt de ces tris successifs est que le tri sur les contextes, faute de très longues répétitions à l'identique, ne crée généralement quasiment pas de doublons, si bien que les clés de tri suivantes sont inopérantes.

2.2. Proposition de généralisation

L'originalité et la force des concordanciers pour l'analyse de corpus, par rapport à de simples relevés d'occurrences, c'est bien cet effet visuel de soulignement des convergences et divergences contextuelles. Cet effet est créé par l'empilement en colonne et les tris, et il est maximisé par un contexte de la taille d'une ligne, sur la largeur d'une page bien dimensionnée pour être embrassée par le regard. Nous proposons donc de nous désintéresser de la possibilité de réglage de la taille du contexte (fonction certes intéressante en soi mais mieux appropriée à d'autres formes de relevés de contextes)[3], et de progresser par la recherche d'un enrichissement des possibilités de tri et d'un renforcement des effets de superposition. Nous proposons donc ici les caractéristiques et fonctionnalités d'une nouvelle génération de concordanciers, sans rompre la lignée des concordanciers maintenant "classiques" qui nous ont directement inspirés.

[2]Le concordancier *Saint-Chef*, mis au point dans le cadre d'un doctorat dédié aux concordances (Sékhraoui, 1995), offrait huit possibilités de tri, certains opérant tour à tour sur un élément du contexte droit et un élément du contexte gauche. Le concordancier *AntConc* (Anthony, 2006) permet de trier sur le *n*-ième mot à gauche ou à droite (pour *n*=1, 2, 3…), avec trois niveaux de tri possibles.

[3]Explicitons peut-être encore ce point difficile : pour un concordancier, enlever le paramètre de réglage de la taille du contexte n'est pas un appauvrissement, selon nous, puisque la variation libre de ce paramètre détruit les effets visuels spécifiques à la concordance. En effet, ces effets de superposition et d'alignement vertical supposent, comme nous l'avons déjà dit, un contexte sur une ligne (et non sur plusieurs), sur la largeur d'une page bien dimensionnée pour être embrassée par le regard. Et nous ne nions pas -bien au contraire- l'intérêt de pouvoir faire varier la taille des contextes extraits d'un corpus (cf. §2.1.2), mais nous pensons que cela relève d'autres fonctionnalités de consultation et d'analyse de corpus (relevé de contextes ou de passages par exemple, cf. (Pincemin, 2006)).

2.2.1. Découpage du pivot en zones

La sélection du pivot se détaille comme une séquence de zones successives : autrement dit, le pivot n'est pas un bloc, mais il est structuré, et composé d'une suite d'éléments individuellement identifiables et potentiellement actifs pour la construction de la présentation des résultats. Concrètement, une interface graphique peut matérialiser cela en proposant une fenêtre pour la désignation du pivot ou de sa première zone, encadrée de boutons permettant l'introduction dynamique, à sa droite ou à gauche, d'une autre fenêtre pour l'indication d'une autre composante du pivot, et ainsi de suite : le pivot se présente alors finalement comme une succession de petites fenêtres, reflétant et détaillant ses zones successives. On peut bien sûr également avoir une version équivalente dans une interface par équation en introduisant, dans le langage de description du pivot, des délimiteurs de zones[4].

2.2.2. Tris et alignements

L'intérêt de ce découpage en zones est de pouvoir indiquer, pour chacune d'elles, (i) si elle est le lieu d'un empilement (en formant une colonne), (ii) si son unité est soulignée par une mise en valeur typographique et laquelle (par exemple couleur, gras), (iii) si elle fait l'objet d'un tri et dans ce cas sur quelle dimension descriptive et de quel type (alphabétique, hiérarchique i.e. par fréquence décroissante, canonique i.e. suivant un ordre conventionnel, cf. Pincemin, 2004). Les contextes gauche et droit disposent également d'une possibilité de tri. Au final le tri de la concordance se définit en fixant un ordre d'application des tris des zones et contextes concernés, s'il y en a effectivement plusieurs.

Les informations (i) et (ii) visent au même effet (faciliter l'alignement, la mise en relation, par le regard) par des voies complémentaires. Elles peuvent ainsi se renforcer (colonne + gras + rouge, par exemple) ; (i) est plus puissant pour les effets d'alignement, mais l'usage de (ii) seul s'avère utile lorsque les occurrences concernées sont potentiellement multiples ou/et à la position très variable[5].

Pour souligner les effets de superposition, la mise en forme de chaque zone correspond à ses caractéristiques de tri (iii) et d'empilement (i). Toute zone triée est alignée à gauche ; en l'absence de tri, un empilement est centré ; une zone ni empilée ni triée est justifiée, et c'est aussi le cas des contextes droit et gauche du pivot en l'absence de tri. En revanche, trié, le contexte gauche est aligné à droite (puisque son tri procède mot par mot de droite à gauche), et le contexte droit, aligné à gauche (comme les zones triées).

2.2.3. Renforcement des atouts des concordances

Un tel concordancier nouvelle génération reprend bien toutes les possibilités de tri imaginées auparavant. Il permet le tri sur des mots distants, tout en maîtrisant mieux la portée des tris[6]. Les effets d'alignement vertical sont démultipliés et mieux caractérisés.

[4]C'est ce que nous avons mis en œuvre dans notre prototype KWIC-LLI, v0.2 (§4).

[5]Dans une variante originale de concordancier, orientée sémantique de corpus, Bourion (2001) met en valeur les mots statistiquement associés au mot pivot : cela se fait naturellement typographiquement (ii) et non par empilement (i), car leur nombre et leur ordre sont *a priori* variables.

[6] Le concordancier *AntConc* (Anthony, 2006), permet de trier sur le n-ième mot à gauche ou à droite du pivot, avec aussi une mise en valeur typographique par des couleurs. Notre découpage du pivot en zones permet un repérage plus fin que la position en nombre de mots par rapport au pivot, puisqu'on peut par exemple s'appuyer sur l'étiquetage ou prendre en compte le contexte.

3. Classes d'objets : de la théorie linguistique à une stratégie d'exploration de corpus

Nous résumons ici les principes linguistiques et méthodologiques pertinents pour la conception du concordancier spécialisé que nous présenterons en dernière partie.

3.1. Prédicats et arguments : de la syntaxe à la sémantique

Les constructions grammaticales s'analysent principalement en relations entre des prédicats (verbes, mais aussi adjectifs ou noms prédicatifs) et leurs arguments (typiquement des noms). Selon une approche de type sémantique distributionnelle, la théorie des classes d'objets ajoute une portée sémantique très forte à cette articulation fondamentale prédicat – argument(s) : un prédicat est défini, non seulement syntaxiquement, mais aussi au plan de son sens, par son schéma d'arguments : un verbe a autant de sens qu'il a de schémas d'arguments. Chaque argument prend sa valeur dans les éléments d'une classe, appelée classe d'objets. Par exemple, le prédicat adjectival *juste* a plusieurs emplois, décrits par sa construction avec différentes classes d'objets : la classe des vêtements (*un pantalon, une veste,... trop juste*), la classe des instruments de musique (*ce piano, cette flûte,... est juste*), etc.

Les prédicats et les arguments se déterminent donc mutuellement au plan sémantique : les classes d'objets regroupent des désignations partageant les mêmes prédicats (par exemple le prédicat *rédiger* délimite la classe des *textes* (*lettre, roman, article*...)). Et réciproquement donc, les sens des prédicats sont déterminés par leur schéma d'arguments, exprimés à l'aide des classes d'objets (comme nous l'avons vu pour *juste*). Une telle description permet donc de définir des classes sémantiques fondées non pas sur une réalité externe ou conceptuelle plus ou moins insaisissable, mais sur une analyse linguistique méthodique.

Les prédicats spécifiques à une classe d'objets sont appelés prédicats appropriés. La délimitation d'une classe d'objets consiste à trouver un ou quelques prédicats appropriés (un faisceau de prédicats appropriés), tel(s) que est élément de la classe tout nom qui peut être argument de (tous) ce(s) prédicats. Par exemple les *vêtements* correspondent aux arguments possibles pour l'ensemble des trois prédicats *mettre* (qqn *met* ...), *être en* (qqn *est en* ...), et *aller bien à* (... *va bien à* qqn).

3.2. Compléter la description par l'observation des contextes en corpus

On considère un corpus étiqueté morpho-syntaxiquement, avec idéalement : identification des unités lexicales, indication de leur catégorie grammaticale (au moins noms, adjectifs qualificatifs et verbes), lemme, dépendances syntaxiques. À défaut des relations syntaxiques, on peut envisager de sélectionner les mots par leur inscription dans une construction linguistique donnée, ou encore (de façon plus ouverte mais moins ciblée) par leur catégorie grammaticale, leur présence dans le voisinage (phrase voire quelques phrases amont pour couvrir des phénomènes d'anaphore) et un indicateur statistique de corrélation (par exemple information mutuelle, écart-réduit ou loi hypergéométrique). Les dépendances entre prédicats et arguments mobilisées par la description en classes d'objets peuvent alors être observées sous quatre angles :

Je cherche à compléter → Je construis des classes de ↓	la caractérisation de l'environnement syntaxique	la composition de la classe
arguments	**Je donne** une liste de noms représentant une classe d'objets. **Je cherche** un faisceau de prédicats appropriés, parmi les verbes, les adjectifs et les noms prédicatifs dont mon objet dépend. **J'indique éventuellement** quelques prédicats pressentis ou déjà retenus.	**Je donne** un ou plusieurs prédicats appropriés (verbe, nom, adjectif). **Je cherche** des noms susceptibles de compléter ma classe d'objets, parmi les noms dépendant syntaxiquement des prédicats indiqués. **J'indique éventuellement** quelques noms pressentis ou déjà éléments de ma classe d'objets.
prédicats	**Je donne** une liste de verbes, d'adjectifs, de noms prédicatifs formant une classe. **Je cherche** des noms ou des classes d'objets à reconnaître comme classes d'arguments, parmi les noms en relation de dépendance syntaxique directe avec le prédicat. **J'indique éventuellement** quelques noms ou classes d'objets pressentis ou déjà retenus, avec pour chacun l'indication de leur position (argument 0, 1, 2, ou 3).	**Je donne** une série de classes d'arguments, par leur composition ou par leur étiquette de classe, en précisant leur place dans la structure d'arguments (0, 1, 2, ou 3). **Je cherche** des prédicats en relation syntaxique avec un tel jeu d'arguments, parmi les verbes, les adjectifs et les noms prédicatifs dans le contexte. **J'indique éventuellement** quelques prédicats pressentis ou déjà retenus.

Pour chaque item indiqué, on peut préciser si on veut le voir rappelé dans les résultats ou non. En effet, donner la description déjà amorcée permet de situer les nouveaux apports par rapport à ce que l'on a déjà. Par la suite, l'indication d'éléments à exclure de l'affichage des résultats rend service soit lorsque la description est déjà très avancée et que ces résultats déjà connus noieraient les nouveaux, soit pour écarter des cas indésirables.

Cette combinatoire équilibrée recouvre en fait des procédures de recherche au statut bien différent. Les cas 1 et 2 (concernant la construction des classes d'arguments) reflètent très directement les principes théoriques, et traduisent donc une exploration très ciblée du corpus. Le cas 3 (partant d'une classe de prédicats, on cherche à compléter sa description syntaxique) correspond à une exploration plus ouverte du corpus : il permet une observation linguistique des contextes (non seulement le schéma d'arguments, mais aussi d'autres propriétés linguistiques pertinentes pour la caractérisation de la classe de prédicats) qui nourrit la description.

En revanche, le cas 4 (complétion d'une classe de prédicats par l'observation des constructions impliquant des classes d'arguments données) est une extrapolation ouverte par cette description combinatoire. C'est une hypothèse demandant à être vérifiée expérimentalement.

4. Un concordancier orienté linguistique distributionnelle

4.1. Principe général : table synthétique

Une recherche des occurrences d'une relation prédicat – argument(s) est lancée sur le corpus, aboutissant à un relevé exhaustif des contextes concernés. Ces contextes, de la longueur d'une

ligne, sont rassemblés sous forme de concordance pour mettre en évidence les régularités de cooccurrence et de construction.

Ce relevé est généralement volumineux. Pour donner une vue d'ensemble, les résultats de la recherche sont présentés de façon synthétique et structurée par un tableau (Figure 2), qui met en évidence les candidats *a priori* les plus intéressants (en vue de compléter la description linguistique). Ce tableau donne ensuite un accès méthodique aux extraits de concordance ciblés (par des liens hypertextes sur chaque case du tableau et dans les marges).

Les colonnes du tableau sont constituées à partir des éléments qui définissent la recherche, et donc par rapport auxquels on peut apprécier les items trouvés dans le corpus. Leur ordre est donné soit par l'utilisateur (pour permettre une organisation mnémonique, significative) soit par les fréquences dans le corpus (pour présenter en premières colonnes les items les plus importants quantitativement).

Les lignes du tableau présentent les résultats de la recherche, par ordre de pertinence décroissante. On distingue graphiquement (par exemple par du gras) les items déjà indiqués au moment de la définition de la recherche. Sont groupés (une seule ligne dans le tableau) les items présentant un profil identique (cooccurrences avec les mêmes têtes de colonne et avec les mêmes fréquences).

Les nombres d'occurrences sont des liens hypertextes vers l'extrait de la concordance correspondant, c'est-à-dire regroupant les cooccurrences des deux items correspondant à la ligne et à la colonne concernées. On a de la même façon des concordances marginales (tous les contextes comprenant l'item correspondant à une ligne, ou tous les contextes correspondant à une tête de colonne) ainsi que la concordance générale, à l'intersection des têtes de ligne et des têtes de colonne. Comme le permettent conventionnellement les navigateurs classiques, lorsqu'un lien a été cliqué il change de couleur, ce qui facilite une exploration systématique de la concordance.

On mentionne également, à côté du tableau,

- les items indiqués et non exclus qui n'ont pas été trouvés dans le corpus : dans certains cas, cette rareté peut conduire à s'interroger sur leur pertinence ;

- les items trouvés mais dans une seule configuration : il peut s'agir de figements ;

- les items trouvés avec une seule occurrence : complètent la recherche, mais peuvent être mêlés de coquilles, ou de propositions trop dispersées.

Ces deux dernières rubriques sont utiles pour alléger le tableau (cf. Figure 3). En effet, la présentation tabulaire n’est réellement pertinente que pour les lignes associées à plusieurs colonnes.

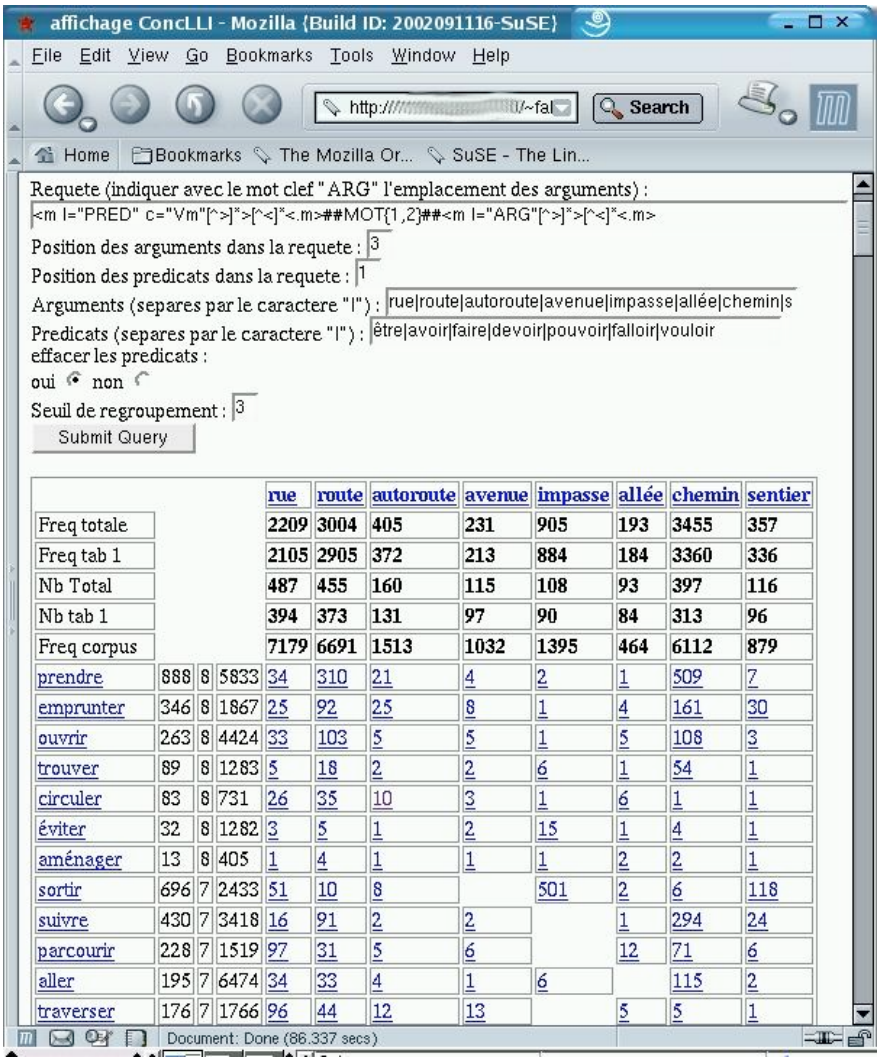
4.2. Illustration : Recherche de prédicats appropriés pour une classe d'objets

Le prototype présenté[7] implémente le premier cas de la combinatoire (§ 3.2.) : on donne une liste de noms représentant une classe d'objets ; on cherche un faisceau de prédicats appropriés dans le contexte de la classe. On peut indiquer quelques prédicats déjà retenus.

[7] Ce prototype est encore en plein développement. Son interface et ses fonctionnalités évoluent. Il n’implémente pas toutes les propositions développées dans la partie théorique, mais il en illustre les principales caractéristiques. La description faite ici ainsi que les copies d’écran correspondent à la version 0.2 du prototype.

Notre exemple utilise un corpus composé de plusieurs années du quotidien *Le Monde*, étiqueté avec l'analyseur morphosyntaxique Cordial. Nous étudions la classe des voies de communication (Mathieu-Colas, 1998).

4.2.1. Tableau synthétique avec accès hypertexte aux extraits de concordance ciblés



				rue	route	autoroute	avenue	impasse	allée	chemin	sentier
Freq totale				2209	3004	405	231	905	193	3455	357
Freq tab 1				2105	2905	372	213	884	184	3360	336
Nb Total				487	455	160	115	108	93	397	116
Nb tab 1				394	373	131	97	90	84	313	96
Freq corpus				7179	6691	1513	1032	1395	464	6112	879
prendre	888	8	5833	34	310	21	4	2	1	509	7
emprunter	346	8	1867	25	92	25	8	1	4	161	30
ouvrir	263	8	4424	33	103	5	5	1	5	108	3
trouver	89	8	1283	5	18	2	2	6	1	54	1
circuler	83	8	731	26	35	10	3	1	6	1	1
éviter	32	8	1282	3	5	1	2	15	1	4	1
aménager	13	8	405	1	4	1	1	1	2	2	1
sortir	696	7	2433	51	10	8		501	2	6	118
suivre	430	7	3418	16	91	2	2		1	294	24
parcourir	228	7	1519	97	31	5	6		12	71	6
aller	195	7	6474	34	33	4	1	6		115	2
traverser	176	7	1766	96	44	12	13		5	5	1

Figure 2 : KWAC-LLI : Accès à la concordance par tableau hypertexte

La concordance calculée est introduite synthétiquement par un tableau avec, en colonne, les éléments de la classe d'objets (lemmes). Les lignes donnent la liste des prédicats (lemmatisés) trouvés.[8] Les prédicats sont triés entre eux d'abord sur le nombre d'arguments différents avec

[8] Les nombres en tête de chaque colonne sont : la fréquence totale du mot (ex. rue) dans les contextes sélectionnés par la requête ; la fréquence des occurrences considérées dans le premier tableau (un second tableau traite des prédicats de basse fréquence en les regroupant) ; le nombre total de prédicats différents recensés dans le tableau et avec lesquels l'argument cooccurre ; le nombre de tels prédicats dans le premier tableau ; la fréquence totale du mot en corpus.
Les nombres en tête de chaque ligne sont : la fréquence du prédicat dans les contextes sélectionnés par la requête (c'est donc aussi la somme des fréquences notées dans les cases) ; le nombre d'arguments (colonnes) avec lesquels le prédicat cooccurre ; la fréquence totale du prédicat en corpus. Les deux premiers nombres sont les deux critères de tri des lignes (nombre d'arguments différents décroissant puis fréquence globale de cooccurrence décroissante).

lesquels ils sont trouvés, puis par leur fréquence totale (ces deux tris sont par ordre décroissant). *Soulignons que ce tri exprime directement un principe théorique très important des classes d'objets : un prédicat a d'autant plus de chance d'être un prédicat approprié définitoire de la classe qu'il est attesté dans des constructions avec un maximum d'éléments de la classe d'objets.* Sont donc présentées en premier les sélections les plus prometteuses.

Il faut cependant avoir écarté les prédicats généraux, attestés eux aussi avec tous les éléments d'une classe. Le plus simple (solution implémentée dans notre prototype v0.2) est d'en déclarer une liste (auxiliaires de temps, modaux, verbe *faire*, etc.) et d'indiquer qu'on exclut ces formes de la recherche. Une heuristique plus ouverte consiste à utiliser une fonction statistique pour trier les résultats (par exemple le critère de l'information mutuelle), de sorte à pénaliser les mots très fréquents qui ne sont pas spécifiquement associés aux mots de la classe. Dans une version ultérieure, notre prototype devrait permettre à l'utilisateur d'activer un tel tri statistique, et même de définir lui-même la fonction à calculer.

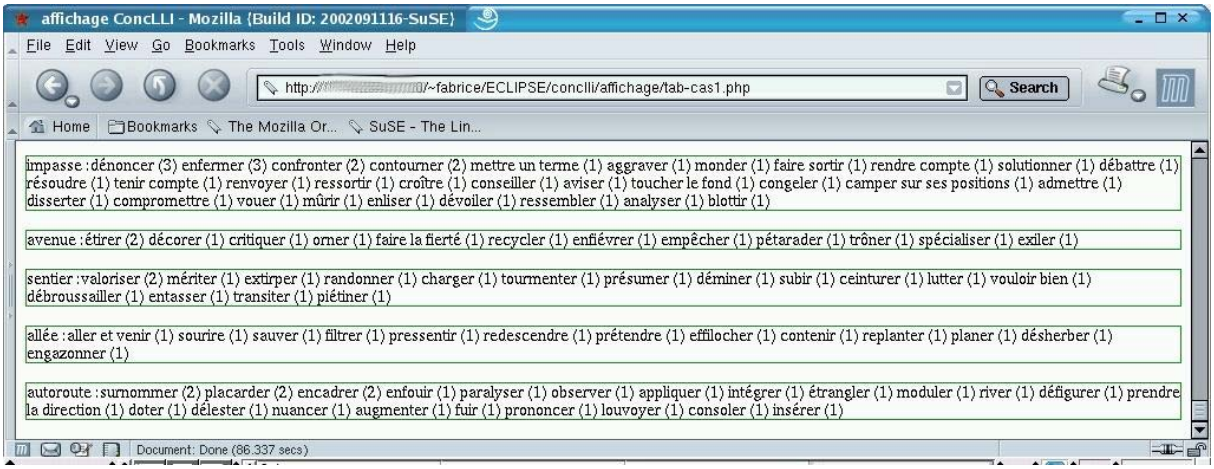
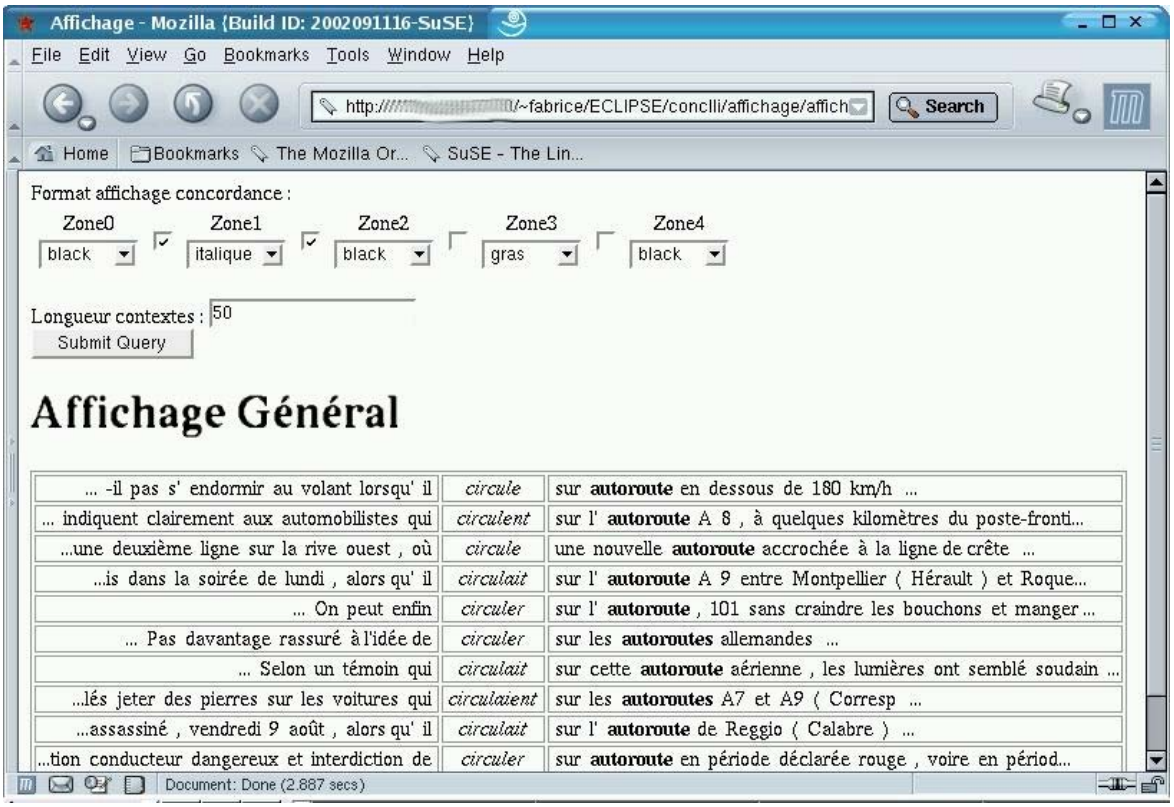
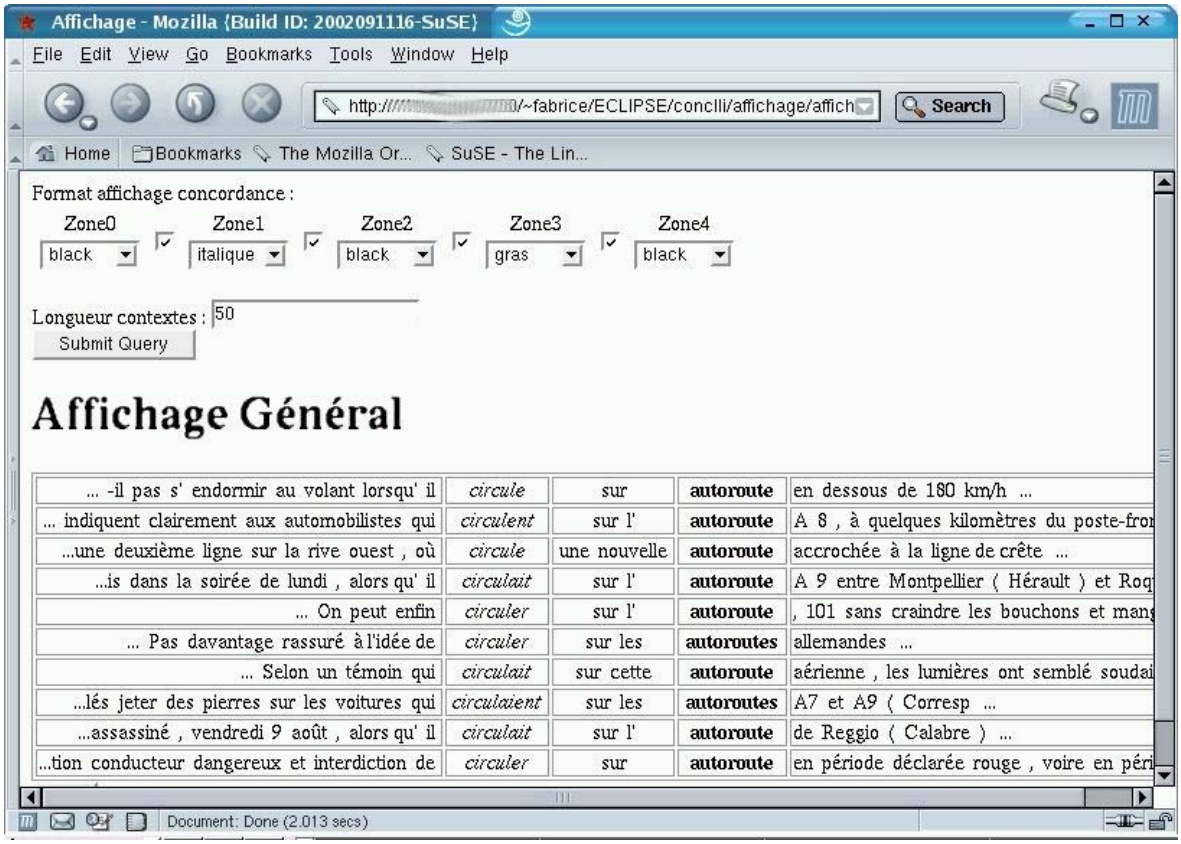


Figure 3 : KWAC-LLI : Après le tableau, les prédicats attestés avec un seul élément de la classe d'objets

4.2.2. Affichage des lignes de concordance en utilisant le découpage du pivot en zones

Le pivot de la recherche (ici la construction dans laquelle cooccurrent le prédicat et l'argument) peut être structuré en plusieurs zones (cf. §2.2.1), grâce au délimiteur ##. L'affichage des concordances est alors paramétrable souplement en indiquant, pour les zones choisies, une mise en forme typographique (couleur, gras, italique, etc.) et (en développement pour une prochaine version) un tri éventuel (dans un premier temps seul le tri alphabétique est proposé). L'effet d'alignement vertical par empilement est obtenu en cochant les points d'empilement voulus, aux frontières des zones.





Figures 4a et 4b : KWAC-LLI : Concordance focalisée sur un couple prédicat-argument (circuler-autoroute) avec marquage des zones du pivot par la typographie et empilement sur une colonne (4a) ou plusieurs colonnes (4b)

5. Conclusion

Bien qu'il s'agisse d'une technique très simple par rapport à des calculs statistiques mis au point en textométrie (Lebart et Salem, 1994), les concordances restent irremplaçables pour une analyse très visuelle et efficace des contextes d'un item linguistique. D'où l'intérêt d'identifier précisément et de développer les caractéristiques qui font leur force, essentiellement les effets d'empilement et les tris mettant en valeur les affinités contextuelles en corpus. Le découpage du pivot en zones est une solution puissante, souple et précise pour mettre en œuvre ces empilements et ces tris.

Le concordancier est alors un outil précieux pour aider le linguiste à dépouiller méthodiquement et rapidement les contextes attestés d'un mot ou d'une classe de mots à décrire. Nous avons présenté ici un logiciel (prototype), KWAC-LLI, particulièrement adapté aux besoins de la théorie des classes d'objets. En particulier, une présentation par tableau hypertexte donne une vue d'ensemble des résultats de la concordance tout en donnant accès au détail des contextes. Le tri des résultats est innovant, il traduit un critère de pertinence linguistique pour l'identification des "prédicats appropriés".

Cette expérience souligne donc l'intérêt toujours actuel des concordanciers tout en proposant des éléments pour l'approfondissement de leurs principes essentiels et le développement de variantes spécialisées.

Références

Anthony L. (2006). *AntConc3.1.2 A Freeware Concordance Program.* Version Windows ou Linux téléchargeable sur http://www.antlab.sci.waseda.ac.jp/

Bourion E. (2001). *L'aide à l'interprétation des textes électroniques*. Thèse de doctorat, Sciences du langage, Université de Nancy II.

Christ O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. *Proc. of COMPLEX'94 (3rd Conf. on Computational Lexicography and Text Research)* : 23-32.

Heiden S. (2002). *Weblex. Manuel Utilisateur*. Version 4.1 (janvier 2002), Laboratoire ICAR, UMR 5191, ENS Lyon.

Le Pesant D. and Mathieu-Colas M. (dir.) (1998). Les classes d'objets, *Langages*, 131.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Mathieu-Colas M. (1998). Les voies de communication. *Langages*, 131.

Mathieu-Colas M. (2005). Les noms de divinités : web, contextes et classes d'objets. *LTT 2005 (7e Journées scientifiques du Réseau Lexicologie, terminologie et traduction) "Mots, termes et contextes"*, Bruxelles, 8-10 septembre 2005.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In Purnelle G. & al., editors, *Proc. of JADT 2004 (7es Journées internationales d'analyse statistique des données textuelles)* : 865-873.

Pincemin B. (2006). Concordances et concordanciers - De l'art du bon KWAC. Soumission à *Documents numériques et interprétation -Corpus en Lettres et Sciences sociales*, Albi, 10-14 juillet 2006.

Sékhraoui M. (1995). *Concordances : Histoire, méthodes et pratique*. Thèse de Doctorat, Université de la Sorbonne nouvelle Paris 3 et École normale supérieure de Fontenay Saint-Cloud.