Le calcul du sens des mots. La lexicologie assistée par ordinateur

Dominique Labbé

Université de Neuchâtel
Institut de Mathématiques
Institut d'Informatique

SEMINAIRE MATHEMATIQUES ET SOCIETE

Mercredi 3 novembre 2010

Ferdinand de Saussure (1857-1913)

Le calcul du sens des mots

La lexicologie assistée par ordinateur

Dominique Labbé
Institut d'Etudes Politiques de Grenoble
Dominique.labbe@iep-grenoble.fr

Résumé : L'étude du langage peut tirer partie des ordinateurs et de la statistique. On présente des nouvelles méthodes qui permettent de calculer le sens des mots chez un auteur, dans un lexique spécialisé ou dans la langue. Ces méthodes peuvent apporter des outils intéressants pour la lexicologie et la lexicographie comme le montre un exemple d'actualité : le sens du mot « banque » dans le vocabulaire économique et social français contemporain.

Abstract : Computers and statistics might be very useful to study languages. We present a new method to measure the contextual meaning of key words in an author's work, in a specialized vocabulary or in a language. This method provides an interesting tool for lexicography as it is shown by applying it to the word "bank" in the French economic and social vocabulary.

Version préliminaire

Il y a 100 ans, en 1910-1911, pour la dernière fois, Ferdinand de Saussure a donné son *Cours de linguistique générale* à l'université de Genève. Ce cours nous est connu grâce à un livre posthume, paru en 1916, rédigé à partir des notes prises par un des étudiants présents. Ce texte a posé les bases de la linguistique et du structuralisme modernes.

Après avoir rappelé les principales propositions de F. de Saussure à propos du sens des mots, on montrera que grâce à la puissance des ordinateurs modernes, ces propositions débouchent sur des applications extrêmement fécondes qui commencent à révolutionner l'étude du langage, notamment la lexicologie (science qui étudie le lexique de la langue), la lexicographie (rédaction des dictionnaires), la terminologie (étude du vocabulaire particulier à une branche du savoir ou de la technique) mais aussi la traduction ou l'histoire littéraire.

Ce sera l'occasion de visiter la bibliothèque du futur. Pour rendre cette visite plus vivante, il fallait un exemple. Puisque, d'après Google, le CreditSuisse.ch est l'institution suisse la plus renommée sur la toile, devant la Confédération et le canton de Vaud (Savoy, 2006), on ne pouvait faire autrement que d'aller à la recherche de la signification du mot *banque* dans le vocabulaire économique français.

I. LA LANGUE ET SON ETUDE

Avant Saussure, on considérait que le sens des mots venait de leur histoire et de leur étymologie. F. de Saussure présente une perspective radicalement nouvelle.

A. Trois caractéristiques de la langue

Voici les trois propositions essentielles de Saussure

1 La langue est un trésor commun à tous ceux qui l'utilisent. Ce trésor est composé de trois systèmes en interaction : une phonétique, une syntaxe, un lexique ;

2 Le signe est l'association arbitraire d'un signifiant (concept) et d'un signifié (assemblage de sons) ;

3 Chaque signe se définit par ses relations avec les autres. Cette proposition s'applique aux trois niveaux définis ci-dessus (phonétique, syntaxe, sémantique).

Enfin, on pourrait ajouter une quatrième proposition : le système est doté des mécanismes nécessaires à son évolution et à son adaptation aux situations concrètes de communication.

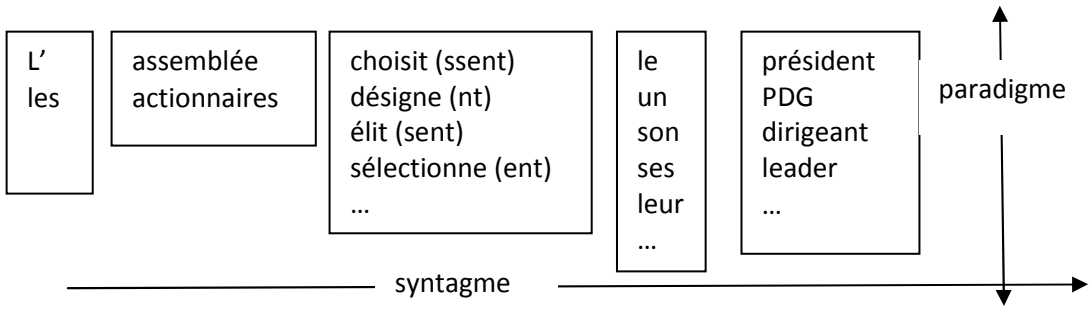
1 Le sens des mots est immotivé…

Commençons par un paradoxe : le signe est arbitraire… et pourtant le locuteur qui l'emploie et le destinataire du message qui le reçoit lui attribuent peu ou prou le même sens.

C'est donc qu'ils partagent ce système de systèmes qui leur permet d'établir le sens d'un mot grâce à un certain nombre d'opérations mentales.

Ces opérations mentales sont d'ordre syntaxique. Par exemple, la catégorie grammaticale du mot, le genre et le nombre des substantifs, la conjugaison du verbe, les règles de composition du groupe verbal, etc.

Elles sont aussi d'ordre lexical : les signes se combinent dans l'énoncé pour former des *syntagmes*, formés de deux ou plusieurs unités consécutives, comme "assemblée générale" ou "élire un président". Les signes s'associent dans l'esprit pour former des *paradigmes* (schéma ci-dessous).



Ces paradigmes sont le produit de trois relations :

Premièrement, *l'hypo et hyperonymie* : relation hiérarchique entre deux vocables dont l'un, de sens plus général, englobe l'autre de sens plus spécifique. Exemples : l'*Etat* englobe le *parlement*, l'*exécutif*... Le *financier* peut être *banquier*, *trader*, *chef d'agence*, *caissier*… ou un simple *usurier*. De ce point de vue, le lexique peut être décrit comme une nomenclature à plusieurs niveaux comme celles qu'utilisent les démographes pour leurs recensements ou les industriels pour standardiser leurs produits.

Deuxième relation paradigmatique : *l'antonymie*, relation d'opposition entre deux vocables. Exemple d'antonymes : noir/blanc ; haut /bas ; grand/petit ; pauvre/riche ; conservateur/libéral, majorité/opposition. Un des deux termes de l'opposition appelle nécessairement l'autre ; pas de mot qui n'ait son antonyme. La recherche du sens d'un mot passe donc par celle de ses antonymes. Et elle aboutit souvent au constat que les deux mots sont étroitement associés dans l'esprit d'un locuteur, comme l'*été* et l'*hiver*.

Troisième relation paradigmatique : la *synonymie* : deux (ou plusieurs) mots partagent au moins en partie le même sens (*chef*, *dirigeant*, *leader*…) Les lexicographes pensent que la synonymie est toujours partielle. Pour reconnaître des synonymes, on recherche les phrases où l'un peut se substituer à l'autre sans changer le sens.

Il faut donc apprendre aux ordinateurs ces relations syntagmatiques et paradigmatiques, c'est-à-dire la syntaxe et le lexique français…

Toutefois, les imperfections de toutes les langues « naturelles » compliquent singulièrement cet apprentissage.

La langue est un système imparfait

Bien que Saussure ne s'en soit guère soucié, il faut rappeler que la langue présente des imperfections, des ambigüités, ce qui se marie mal avec l'informatique ! Signalons notamment :

- la *polysémie* est certainement le défaut le plus évident. Comme la suite de cette conférence le montrera, le substantif féminin *banque* est fortement polysémique, c'est-à-dire que plusieurs concepts sont associés à un même signifiant. En fait, tout le vocabulaire usuel est sujet à cette polysémie. Le vocabulaire spécialisé est en grande partie constitué pour lutter contre ce défaut et pour tenter d'assurer l'unicité signifiant/signifié (Favre et al 1997) ;
- l'*homonymie* : plusieurs signifiés différents et plusieurs signifiants mais une seule graphie : *politique* est à la fois substantif masculin, substantif féminin et adjectif…
- l'*homographie* : deux signifiants différents mais de même graphie. Par exemple "été" (substantif masculin et participe passé du verbe être) ; "est", point cardinal et troisième personne de l'indicatif présent du verbe "être"), finance (substantif féminin et verbe *financer*) ;

- l'*homophonie* : deux signifiants différents ont une même prononciation malgré une orthographe différente. Par exemple, "est" (troisième personne de l'indicatif présent du verbe être) et la conjonction "et", se prononcent pareil et s'écrivent différemment.

2. L'atelier du lexicographe

Les grammaires et les dictionnaires ne donnent que des images imparfaites de la syntaxe et de la sémantique de la langue telle qu'elle est pratiquée par ses usagers. Comment l'informatique et la statistique peuvent-elles améliorer ces outils ? Comment des automates peuvent-ils utiliser les principes posés par Saussure ?

Pénétrons dans l'atelier du lexicographe et voyons quels sont ses procédés, du moins quand il est débarrassé de l'illusion selon laquelle le sens d'un mot lui vient de son histoire.

La démarche du lexicographe - pour définir le(s) sens d'un mot chez un groupe d'usagers ou dans l'ensemble de la communauté parlant une langue - se déroule en trois temps (par exemple : Blumenthal & Hausmann 2006).

D'abord, il recherche des citations dans la littérature ou la presse et, quand il n'a rien de satisfaisant, il forge des exemples théoriques (« exemplier »). Par intuition et approximations successives, il recherche les meilleures paraphrases possibles de ces citations et exemples.

Jusqu'à maintenant, le lexicographe a fait ces recherches manuellement – ou avec une assistance informatique très limitée - sans avoir la certitude de ne pas passer à côté de certains usages et en hiérarchisant les différents emplois de manière intuitive.

Ceci fait, il recense les **syntagmes** que l'on peut former avec ce mot. Par exemple, "banque d'affaires" ou "banque de dépôt" sont des syntagmes. Enfin il recherche les mots qui peuvent se substituer au mot étudié dans certaines de ces paraphrases. L'ensemble de ces substituts forme le **paradigme** du mot recherché.

Enfin, le lexicographe en tire un "article de dictionnaire", censé restituer le mieux possible le(s) sens du mot dans la langue. Sous l'entrée (mot vedette et catégorie grammaticale), l'article énumère les différents sens possibles du mot en donnant, pour chacun de ces sens, une définition, des synonymes, des antonymes et des citations illustratives.

Deux remarques.

Premièrement, nous montrerons que certaines de ces opérations peuvent être avantageusement prises en charge par les ordinateurs.

Deuxièmement, malgré le talent des lexicographes français et la qualité formelle des dictionnaires, le résultat semble parfois décevant ou peu en phase avec la langue telle qu'on la parle et qu'on l'écrit effectivement… Pourquoi cette insatisfaction ?

B. Les difficultés de la lexicographie française

Rappelons l'idée essentielle de Saussure : une langue est un trésor commun à tous les usagers de cette langue. Connaître une langue, c'est observer les usages de celle-ci. Pourtant les linguistes et les grammairiens français ne le font guère, non pas par mauvaise volonté – ou par ignorance - mais parce qu'il leur manque les outils nécessaires pour cette observation, c'est-à-dire les enquêtes d'usage et les grands corpus étiquetés.

1. Absence d'enquête d'usage

La seule enquête scientifique sur l'usage du français date de plus d'un demi-siècle. Elle a été pilotée par G. Gougenheim (1900-1972). Au début des années 1950, lui et sa petite équipe ont enregistré des locuteurs de tous les milieux, à propos de leur vie quotidienne, de leur travail de leurs loisirs…, puis ils ont saisi ces enregistrements et en ont réalisé un traitement statistique simple qui a abouti à une grammaire élémentaire du français et à un vocabulaire fondamental contenant les 3 500 mots les plus utilisés du français avec des phrases canoniques (Gougenheim, 1956 et 1958). Avec les moyens de l'époque, c'était un travail pionnier remarquable. Il a été vivement critiqué : les enquêteurs avaient enregistré des gens ordinaires et non pas les "spécialistes" : intellectuels, linguistes ou grammairiens…

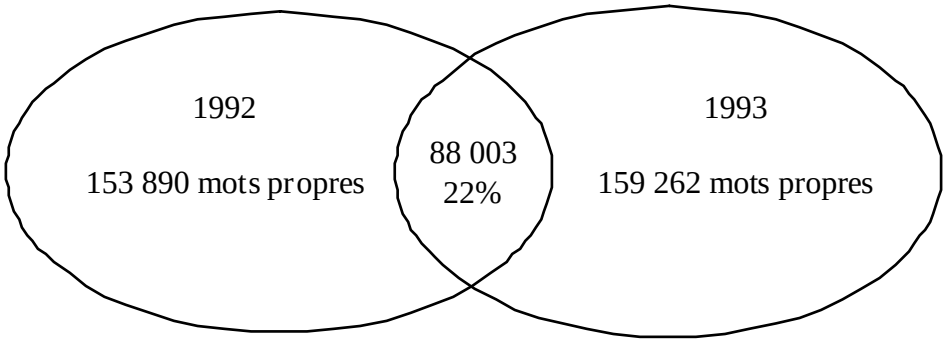
On entend souvent dire que cette absence d'enquête d'usage n'est pas grave puisque, aujourd'hui sur la toile, on dispose d'une gigantesque collection de textes français qui serait une source inépuisable d'informations sur notre langue. Pourtant, les lexicographes n'utilisent pas beaucoup ces grandes bases textuelles, notamment les grandes bases en ligne comme « Google livres". Quelques audacieux y trouvent de quoi renouveler leur stock de citations illustratives. C'est tout. Pourquoi ?

D'abord parce que rien ne peut remplacer une enquête menée dans les règles de l'art sur la manière dont parle (et non pas écrit) un échantillon représentatif de la population. Et deuxièmement, parce que les grandes collections de textes électroniques sont très imparfaites. On sait bien que les moteurs de recherche, qui permettent de les consulter, ne sont pas exempts de reproches (Savoy 2009). De plus, elles sont pratiquement impossibles à utiliser pour une étude du lexique du français.

2. L'apport limité des collections électroniques de textes

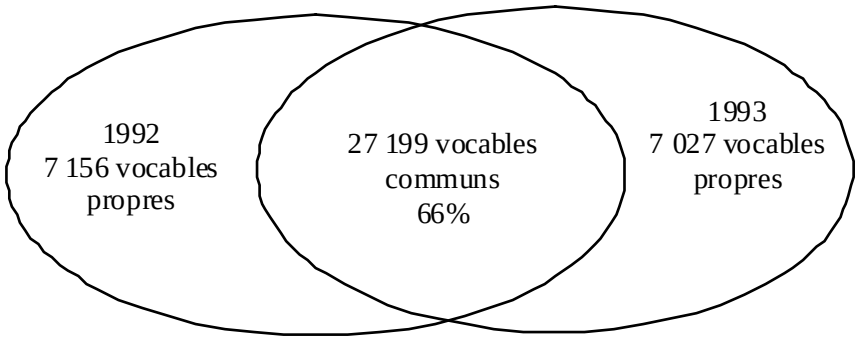
Les collections électroniques des grands journaux offrent apparemment un l'outil idéal. En fait, c'est une illusion. Voici une expérience réalisée par Silberztein (1995) sur deux années du journal *Le Monde* (1992-1993). L'expérience porte sur un total 45 millions de mots qui se répartissent ainsi.

Le Monde	1992	1993	1992-93
Occurrences (mots)	21 804 745	23 198 877	45 003 622
Formes graphiques (mots différents)	241 893	247 265	401 155



Le "vocabulaire" de ces deux années compte au total plus de 400.000 "mots" différents et comporte un noyau commun très petit. 78% du lexique du français aurait-il changé d'une année sur l'autre ? Le journal aurait-il traité de sujets totalement différents entre 1992 et 1993 ?

En fait, la plupart des "mots différents" sont des fantômes : coquilles ou fluctuations orthographiques – notamment sur les noms propres – diverses conjugaisons d'un même verbe, etc. Après correction orthographique, standardisation des graphies et "lemmatisation", le tableau est radicalement différent.



Le vocabulaire ne compte plus "que" 41.000 entrées dont les deux tiers sont communes aux deux années…

A l'époque, *Le Monde* disposait d'une armée de correcteurs très vigilants. La qualité était donc la meilleure possible. Toute autre expérience, sur la toile notamment, donnera des résultats encore plus décevants.

Actuellement, les grandes collections électroniques de textes sont des "émeutes de formes", comme "le pittoresque est une émeute de détails" (Baudelaire). Trois opérations préalables sont indispensables pour rendre exploitables ces collections électroniques de textes : correction orthographique, standardisation des graphies – spécialement des noms propres - et "lemmatisation".

A ce prix, on peut disposer de grandes bases de données lexicales dont le modèle est le British National Corpus (hébergé par l'Université d'Oxford). Ce corpus de 100 millions de mots est représentatif des usages de l'anglais contemporain. Il comporte une section dédiée à l'anglais parlé, élaborée grâce à une enquête d'usage sur un échantillon représentatif de la population du Royaume Uni (Burnard, 1995 Crowdy 1993, Nelson 1997).

De tels corpus sont indispensables pour passer d'une lexicographie artisanale à une lexicométrie scientifique.

C. De la lexicographie à la lexicométrie

La lexicométrie marie les outils traditionnels de la lexicologie – tels qu'ils sont hérités de Saussure - avec la science moderne, notamment l'informatique et la statistique.

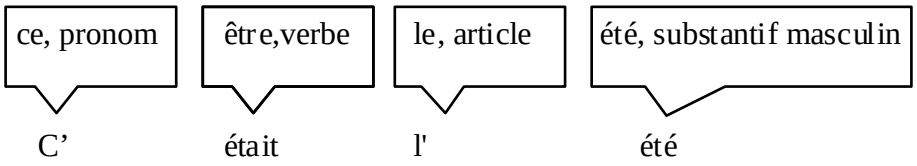
1. La bibliothèque du futur

La bibliothèque du futur sera électronique. Son organisation ne sera guère différente de celle des bibliothèques actuelles. Elle aura des catalogues où seront enregistrés les ouvrages et les articles qu'elle rangera dans des dossiers équivalents aux rayonnages sur lesquels on range les livres. Mais elle présentera quelques avantages dont les plus évidents seront certainement qu'on accédera immédiatement aux documents recherchés – s'ils figurent au catalogue - qu'elle ne sera jamais fermée et qu'elle sera consultable à distance.

Pour entrer dans cette bibliothèque, chaque texte devra subir un certain nombre de traitements préalables avant son indexation.

Les traitements préalables.

La constitution de catalogues et d'index sont évidentes. Mais le texte lui-même doit être retravaillé. Nous avons déjà signalé la correction orthographique et la standardisation des graphies. Puis le texte est découpé en autant d'emplacements (en anglais "tokens") qu'il y a de mots et chacun de ces "mots" est doté d'une étiquette. On désigne cette opération sous le nom de "lemmatisation" ("étiquetage" serait préférable). En voici un exemple :



L'étiquette attachée à chaque mot du texte, comporte son "entrée de dictionnaire" et sa catégorie grammaticale. Par exemple, la forme "été" peut recevoir deux étiquettes : "être, verbe au participe passé" ou "été, substantif masculin". Ou encore "l'" : "le, déterminant" ou "le, pronom".

Cette opération est indispensable. Elle a beaucoup de justifications pratiques. Par exemple, pour le français, c'est le seul moyen de retrouver les verbes qui seront rattachés à leur seul infinitif alors qu'ils ont parfois jusqu'à cinquante flexions différentes. C'est surtout le moyen de départager les homographes (une même graphie mais deux sens différents, comme *été* ou *le*). Dans un texte en français, en moyenne un tiers des mots sont des homographes.

Cette opération a aussi une justification théorique évidente. Comme indiqué en introduction, chaque mot tire son sens de sa place dans la langue. La lemmatisation consiste à replacer chacun des "mots" du texte à sa place dans le système de la langue, exactement comme le fait le lecteur. Ce sera le seul moyen d'accéder au sens précis que ce mot prend dans la langue aussi bien que dans une communauté particulière ou un auteur particulier.

Naturellement, le texte original n'est en rien modifié. On y ajoute des étiquettes qui sont autant de portes d'entrée dans le texte, comparable aux entrées d'un dictionnaire.

En voici un exemple tiré de notre bibliothèque électronique qui comporte actuellement 18 millions de mots étiquetés dont près de 8 millions sont tirés d'œuvres littéraires françaises de ces 4 derniers siècles. Naturellement, cette bibliothèque ne peut être considérée comme un échantillon représentatif du français moderne mais comme une simple esquisse.

L'été chez le Clézio

JM Le Clézio (1940-) a reçu le prix Nobel de littérature en 2009.

Dans notre bibliothèque électronique, le corpus Le Clézio comprend toutes les œuvres parues entre 1963 (son premier roman *Le procès verbal*) et 1999 (le *Hasard*), soit une douzaine de livres (dont deux nouvelles et trois recueils de nouvelles) comportant au total 870 467 mots et un vocabulaire de 18 001 vocables différents.

Le vocabulaire de Le Clézio – comparé à celui des autres romanciers français – met en valeur quelques substantifs dont "œil", "mer"… "été". L'*été* est sa saison préférée (on le sait grâce à un test statistique que nous allons présenter dans la suite de cette conférence).

Une interrogation de la bibliothèque permet d'obtenir immédiatement la liste des 123 emplois de ce substantif. Voici le début de cette "concordance" pour le roman "Désert".

est arrivée. Il faisait très chaud parce que c'était l'	été	, et le vent soulevait des nuages de poussière
devais naître est arrivé, c'était peu de temps avant l'	été	, avant la sécheresse. Hawa a senti que tu allais
aussi, l'odeur de la mer et du vent, des prairies en	été	. Il y a tout cela, et bien davantage, dans cette pl
artani allume les étoiles, une, une, encore une… L'	été	, la pluie commence à tomber, l'eau coule dans
ssi. Quand il commence à pleuvoir, au milieu de l'	été	, l'eau ruisselle sur les toits de tôle et de papier
La maison des bains ne fonctionne que pendant l'	été	, parce que l'eau est rare, ici. L'eau vient d'une
les jours de pluie, les jours de vent, les jours de l'	été	. Quelquefois Lalla croit qu'elle attend seulem
aison d'Aamma, un matin, au commencement de l'	été	. C'était un homme de la ville, habillé avec un
le mendiant qui lui a montré où il passe les nuits, l'	été	, quand le vent qui vient de la mer est tiède com
et dans les huttes de branches. Le vent chaud de l'	été	les couvrait de poussière, mais ils attendaient,
brillante sous le bleu du ciel. Le vent ardent de l'	été	passait sur la terre, soulevait la poussière, voila
déjà très bleu, limpide, sans un nuage. Le vent de l'	été	souffle de la mer, s'engouffre dans les rues, le lo
es grands araucarias. Radicz aime bien le vent de l'	été	; ce n'est pas un vent mauvais, comme celui qui
est réveillé et il a senti tout de suite que le vent de l'	été	avait commencé. Alors il s'est un peu roulé dans
habits, et de plonger dans l'eau. C'est le vent de l'	été	qui l'a appelé jusqu'à la mer, qui lui a montré l'e
ncore baissés, les balcons sont vides. Le vent de l'	été	souffle sur la façade des immeubles et fait claq
, comme venu d'un autre monde, le vent de l'	été	a endormi tous les habitants et toutes les habita

La concordance complète montre l'association de l'*été* avec le *vent*, la *chaleur*, le *soleil*, le *ciel* et la *mer* qui sont des thèmes très présents dans l'œuvre de cet écrivain. Naturellement, pour une analyse approfondie, on peut élargir le contexte, afficher des paragraphes entiers, etc.

Deux remarques :

Les textes de Le Clézio comptent au total 870 470 mots dont 123 substantifs "été" (et 122 au singulier), contre 548 "été", verbe être au participe passé. Dans une collection traditionnelle de textes, le lecteur aurait dû chercher les 122 substantifs au milieu des 548 participes, sans être certain de les trouver tous. S'il parvient au bout de cette tâche, il lui sera impossible de savoir que l'*été* est un mot favori de Le Clézio, sauf à faire le même travail sur les autres romanciers…

Cette concordance sur les lemmes n'est pas disponible pour les chercheurs français, car aucune des collections de textes français consultables sur internet n'est étiquetée.

Deuxièmement, comment peut-on affirmer que, dans l'esprit de J.-M Le Clézio, l'*été* est associé avec quelques autres thèmes comme le *vent* et la *mer* ?

La statistique lexicale permet de répondre à ces questions.

2. La statistique lexicale

On cherche le sens que J.-M. Le Clézio donne au vocable "été". On demande à l'ordinateur d'effectuer les opérations suivantes :

- sortir de la bibliothèque tous les ouvrages disponibles de Le Clézio. Ils constitueront le "corpus",
- relever, dans ce corpus, toutes les phrases qui contiennent ce vocable. Cet ensemble sera l'"univers" de l'*été* chez Le Clézio,
- établir le vocabulaire de ce sous-ensemble,
- comparer ce vocabulaire à celui des autres phrases de l'auteur.

Quand Le Clézio parle de l'*été*, quels sont les mots qui sont trop employés ou pas assez (par rapport au reste de l'œuvre) ?

- Les premiers sont positivement associés ou encore, il y a une relation d'attirance : quand l'auteur pense à l'un, l'autre lui vient à l'esprit.
- Les seconds sont négativement associés : relation de répulsion mutuelle.

Le sens du vocable – "été, substantif masculin" – chez Le Clézio est l'ensemble du vocabulaire sur-employé ou sous-employé autour de ce vocable. Cette liste ne résulte pas des intuitions du critique littéraire mais d'un calcul. Ce calcul a été présenté pour la première fois dans Labbé 1994 et il a fait l'objet d'une première publication scientifique dans : Labbé et Labbé 2005.

Voici ce calcul appliqué à la liaison entre deux vocables "été" et "mer".

Notons :

C. l'ensemble du corpus, ici les œuvres de Le Clézio entre 1965 et 1999.

N_c : la longueur du corpus : nombre de mots dans le corpus (soit 870 467 mots)

U : les phrases contenant le vocable recherché (*été*)

N_u : le nombre de mots dans les phrases contenant le vocable recherché (ici 3 007 mots)

C-U. les phrases ne contenant pas *été*.

F_{ic} : le nombre des occurrences du vocable *i* dans l'ensemble du corpus (ici *mer* : 1857)

F_{iu} : le nombre des occurrences du vocable *i* dans U (ici 15)

E_{iu} : l'espérance mathématique du vocable *i* dans U :

$$(1)\ E_{iu} = F_{ic} * \frac{N_u}{N_c} = 1857 \frac{3007}{870\,467} = 6.4$$

E_{iu} peut se lire ainsi : si le vocabulaire de Le Clézio était uniformément réparti sur l'ensemble de son oeuvre, les vocables *mer* et *été* seraient associés 6 à 7 fois au sein d'une même phrase.

La fréquence constatée ($F_u = 15$) est supérieure à cette fréquence attendue. Peut-on dire que, dans l'œuvre de Le Clézio, il existe une relation d'association entre ces deux vocables ?

Cette relation peut se mesurer à la probabilité *P(X)* de l'événement observé F_{iu} par rapport à l'événement attendu (E_{iu}). Cette probabilité est le produit de deux événements :

- le choix de N_u objets parmi N_c :

$$C_c^u = \frac{Nc!}{Nu!(Nc - Nu)!} = \begin{bmatrix} Nc \\ Nu \end{bmatrix}$$

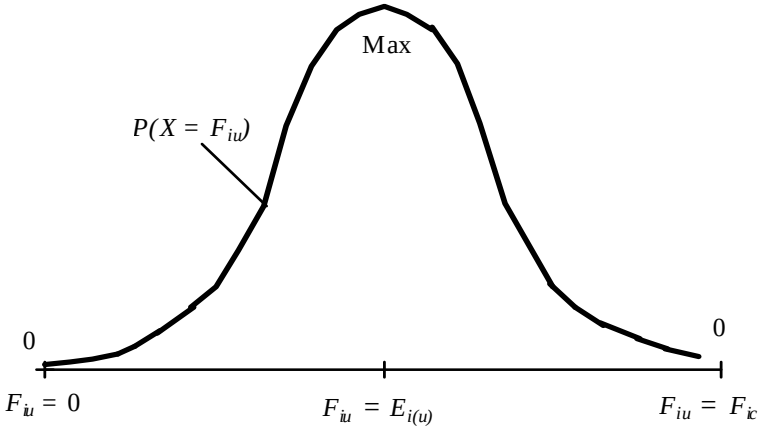
- le choix de F_{iu} parmi F_{ic} objets :

$$C_{Fic}^{Fiu} = \frac{Fic!}{Fiu!\,(Fic - Fiu)!} = \begin{bmatrix} Fic \\ Fiu \end{bmatrix}$$

La probabilité que ces deux événements surviennent concurremment suit une loi hypergéométrique de paramètres F_{ic}, F_{iu}, N_u, N_c:

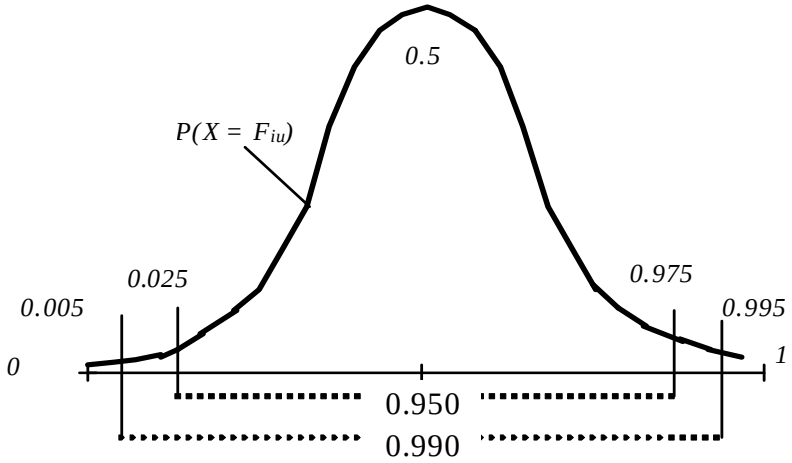
$$(2)\ P(X = F_{iu}) = \frac{\begin{bmatrix} F_{ic} \\ F_{iu} \end{bmatrix} \begin{bmatrix} N_c - F_{ic} \\ N_{iu} - F_{iu} \end{bmatrix}}{\begin{bmatrix} N_c \\ N_u \end{bmatrix}}$$

A condition que N_u et F_{ic} soient suffisamment grands, les valeurs de la variable X sont distribuées le long de l'intervalle [0-F_{ic}], selon une courbe dont le mode (probabilité maximale) est atteint quand $E_{iu} = F_{ic}$ (figure ci-dessous) :



$F_{iu} = 0$ aucune occurrence du vocable *i* dans l'univers étudié,
$F_{iu} = E_{iu}$ la fréquence observée correspond à l'espérance mathématique de la variable,
$F_{iu} = F_{ic}$ toutes les occurrences de *i* se trouvent dans l'univers et aucune dans le reste du corpus (C-U).

Soit L_{iu} un indice permettant de répondre par un seul chiffre à deux questions : y-a-t-il un lien entre le vocable *i* et l'univers considéré ? Si oui, de quelle intensité et de quel sens (attirance ou répulsion) ? On propose d'utiliser les intervalles usuels en calcul de probabilité :



Si l'on désire réduire le calcul aux liens les plus sûrs, on retiendra les vocables pour lesquels $L_{iu} < 0.005$ ou $L_{iu} > 0.995$. En revanche, si l'on souhaite un recensement plus large mais moins assuré, on retiendra les bornes $L_{iu} < 0.025$ ou $L_{iu} > 0.975$. Dans ce dernier cas, on devra considérer avec précaution les mots les plus proches des bornes de l'intervalle.

Concrètement, l'indice sera le cumul des valeurs que peut prendre la variable, en partant de 0 et en allant jusqu'à F_{iu}

$$(3)\ L_{iu} = P(X \leq F_{iu}) = \sum_{j=0}^{j=F_{iu}} P(X = j)$$

Dans le cas de Le Clézio, l'indice de liaison entre *mer* et *été* est égal à .991. On accepte cette liaison au seuil de 5% mais non au seuil de 1%. Le principal intérêt du calcul est de valider (ou d'invalider) les intuitions suggérées par les concordances. Par exemple, chez Le Clézio, la liste des substantifs liés à l'*été* (au seuil de 5%) est la suivante (classée par ordre décroissant de l'indice de liaison :

vent, chaleur, hiver, robe, pluie, mer, jour, eau

Seuls les 5 premiers ont un indice supérieur à .995. On ne sera pas surpris par le "vent d'été" ou la "chaleur (de l') été", ou une "robe d'été". En revanche, il est intéressant de confirmer que "mer" est bien associée à "été" dans l'esprit de Le Clézio ; il est encore plus intéressant de constater l'association de "hiver" avec "été", association qui serait certainement passée inaperçue à la lecture des concordances. Enfin, les trois vocables les plus "sous-employés" avec "été" sont : "œil", "voir" et "savoir". Cela serait sûrement passé inaperçu à la lecture des concordances car comment y repérer les vocables qui n'y sont pas mais qui devraient s'y trouver ? Et comment savoir que le regard (*œil* et *voir*) sont les thèmes qui singularisent Clézio ? Que les autres thèmes propres à Le Clézio sont la *lumière* (*soleil*), *terre* et le *ciel*, l'*eau* et la *mer*, la *nuit*, etc… Il faut pour cela dépouiller le reste de la littérature avec les mêmes automates…

Deux remarques de méthode :

Premièrement, le calcul n'a de sens que lorsque F_{iu} ou E_{iu} sont suffisamment grandes pour que les seuils indiqués ci-dessus puissent être atteints. Par exemple, cela signifie que si E_{iu} est inférieure à 0.5, F_{iu} doit être au moins égal à 5 et que si F_{iu} est nulle, E_{iu} doit au moins être égale à 5. En pratique, cela limite le calcul aux vocables les plus fréquents. Dans le cas contraire, l'absence de lien peut simplement provenir de ce que l'un des deux vocables a un nombre d'occurrences insuffisant pour entrer dans le calcul ;

Deuxièmement, la formule (3) a deux limites : $F_{ic} < N_U$ et $F_{ic} < (N_c - N_U)$. La borne supérieure signifie que le calcul est inutile en cas de corpus monothématique (la plupart des phrases appartiennent à U). La borne inférieure signifie que le calcul doit porter sur de grands univers, ou que, pour des petits univers, on doit exclure les vocables les plus fréquents (les "mots-outils"). En pratique, cette restriction concerne les prépositions et articles les plus usuels ainsi que les verbes *être* et *avoir*…

L'outil permet d'aller encore beaucoup plus loin, comme nous allons maintenant le montrer grâce à notre bibliothèque électronique et au mot "banque".

II. UNE APPLICATION : LE SENS DU MOT *BANQUE*

Pour illustrer cette méthode, nous prendrons l'exemple du substantif féminin "banque" dans le vocabulaire de l'information économique contemporaine.

Auparavant, on pourra consulter un dictionnaire de langue ou de synonymes. Par exemple, dans le dictionnaire *Robert des synonymes* : "caisse de crédit, de dépôts, établissement de crédit, comptoir". Cette courte liste correspond-elle aux emplois des locuteurs et notamment des acteurs de l'économie ?

Pour répondre à cette question, on utilise une autre partie de la bibliothèque électronique : 1 064 articles parus dans les rubriques économiques de cinq journaux (*Les Echos*, *Le Monde*, *Le Nouvel économiste*, *Capital*, l'*Expansion*) entre janvier 1996 et décembre 1998, soit 1,4 millions de mots (environ 300.000 pour chacun des journaux), et 33 034 vocables. Ce corpus a été constitué avec J. Leselbaum (société Signifier) à la fin des années 1990 pour étudier le vocabulaire de la presse. Une première version de ce travail a été présentée en février 2002.

Au 15e rang des substantifs les plus employés, on trouve *banque* avec 1 233 occurrences.

Les opérations statistiques décrites ci-dessus débouchent sur les conclusions suivantes.

A. Un univers singulier

1. Une entité abstraite mais présente dans tous les sujets.

Le corpus comporte 62 231 phrases, soit une longueur moyenne des phrases de : 23,35 mots. Il y a 1 140 phrases contenant le mot *banque*. Elles comportent au total : 35 145 mots, soit une longueur de 30.83 mots par phrases.

La longueur des phrases contenant le mot *banque* est donc particulièrement élevée.

Il s'agit d'un mécanisme psychologique sur lequel on reviendra plus loin : quand un sujet est particulièrement important, ou complexe, on a tendance à faire des phrases plus longues et plus compliquées qu'à l'ordinaire et à choisir un degré d'abstraction plus élevé.

Deuxième caractéristique intéressante, sur les 274 vocables spécifiques à l'univers de *banque*, 180 – soit les deux tiers - sont des associations positives ("trop utilisés") et seulement un tiers (94) des associations négatives (antonymes). Dès qu'un mot est beaucoup employé (comme c'est le cas de *banque*), on observe habituellement un nombre à peu près équivalent d'associations positives et négatives. Ici le déséquilibre est très important. Cela signifie qu'il y a peu de domaines (et de thèmes) dont les banques soient absentes ou peu impliquées… mais surtout que, dans l'esprit des observateurs, la notion même de banque n'est pas très précise puisqu'*ils ne savent pas très bien ce que les banques ne sont pas et ce qu'elles ne font pas* !

2. Un univers orienté vers le nom

Le tableau en annexe 1 présente les densités des catégories grammaticales utilisées dans le reste du corpus comparées à ces mêmes densités dans l'univers de banque. On constate :

- un recul important du verbe – spécialement toutes les formes de l'indicatif (-10%) - ce qui est considérable. Le recul le plus important concerne les pronoms personnels (-27.6% : les propos sur les banques sont dépersonnalisés) et les adverbes (-9,2) ;
- un gonflement en sens inverse du groupe nominal, spécialement les noms propres, les adjectifs et leurs satellites (articles et prépositions).

En français, les verbes ont des densités d'emploi reliées à celles des adverbes et des pronoms, spécialement les pronoms personnels qui amplifient toujours les mouvements du verbe. Le poids relatif de ce groupe est inverse à celui des substantifs, des adjectifs et des déterminants. Autrement dit, les groupes nominal et verbal sont opposés. Le groupe verbal est constitué des pronoms, des verbes, des adverbes et des conjonctions de subordination ; le groupe nominal est constitué des substantifs, adjectifs, déterminants, prépositions et conjonctions de coordination. Certes, le partage n'est pas exclusif : certains adverbes peuvent se glisser dans le groupe nominal, certaines prépositions dans le groupe verbal, certains pronoms relatifs s'utilisent dans les deux, etc. Enfin les locutions, les mots étrangers, etc. ne sont pas classables dans l'un ou l'autre des groupes. Ces réserves admises, voici les proportions observées dans l'ensemble des textes écrits de la bibliothèque ("français général", dans la presse économique et dans l'univers de *banque*.

	Français général	Presse économique	Univers de *banque*
Groupe verbal	36%	23%	20%
Groupe nominal	64%	77%	80%

En moyenne, dans les textes écrits en français depuis le XVIIe siècle – et présents dans la bibliothèque électronique -, le groupe verbal couvre 36% de la surface des textes contre 64% pour le groupe nominal (en négligeant les locutions et mots étrangers). Dans le corpus presse économique entier, le groupe verbal ne couvre que 23% de la surface contre 77% au groupe nominal. Ces proportions passent à 20% et 80%, quand il est question des banques.

Les journalistes économiques privilégient donc manifestement le groupe nominal, spécialement l'adjectif et le discours impersonnel (très fort déficit en pronoms). Quand ils traitent de sujets touchant aux banques, ils aggravent encore cette caractéristique.

Comment interpréter la préférence pour le groupe nominal et la "fuite" devant le verbe ?

Plusieurs interprétations complémentaires sont proposées.

Pour la stylistique traditionnelle, la construction nominale "présente le fait sans date, sans mode, peut-être sans aspect, sans le rattacher nécessairement à un sujet (donc à une cause), à un objet (donc à un but)" (Cressot 1963, p. 154). Cela peut sembler contradictoire avec le fait que la plupart des articles constituant le corpus sont surtout des reportages…

Pour la linguistique, le verbe (ou ses équivalents) a une double fonction : la "fonction cohésive" qui organise "en une structure complète les éléments de l'énoncé" et la fonction assertive qui "dote l'énoncé d'un prédicat de réalité" car l'élément verbal implique une référence à un ordre qui n'est plus simplement celui du discours mais celui de la réalité (Benveniste 1981, 1, p. 154). Si l'on adopte ces hypothèses, la préférence pour le groupe nominal permettrait aux journalistes économiques d'effacer de leurs textes (au moins partiellement) les questions pour lesquelles ils n'ont pas la réponse ou qui leur semblent hors de portée.

Troisièmement, en 1950, le statisticien Guiraud avait signalé que le nombre des substantifs et celui des verbes varient en sens inverse et que le substantif domine dans la "prose abstraite". Il a été également montré comment, chez le même auteur, le passage de l'oral à l'écrit se traduit par une diminution très significative du poids du groupe verbal et une augmentation parallèle du groupe nominal. Autrement dit, l'expression spontanée privilégie le verbe, les pronoms, les adverbes. Le passage à l'écrit amène à remplacer un certain nombre de ces verbes par des substantifs, certains adverbes par des adjectifs, à réduire l'emploi du démonstratif, etc. A la suite de Guiraud, on peut donc penser que l'effort d'élaboration qu'implique l'écrit s'accompagne d'un mouvement d'abstraction au-delà de la perception ou de la visée immédiate de l'auteur.

De ce point de vue, le constat le plus significatif concerne les chiffres. Au centre de l'univers des banques, on s'attendrait à trouver aussi les monnaies et les chiffres, or c'est le contraire qui se produit : l'emploi des chiffres est de -16.8% inférieur à ce qu'il est dans le reste du corpus. Dans la liste des substantifs significativement sous-employés (annexe 2), on trouve : *milliard*, *franc*, *dollar* ainsi que tous les chiffres (dans les déterminants)... Les chiffres sont le principal ancrage dans la réalité économique. Leur faible emploi confirme le passage à l'abstraction lorsqu'il est question des banques.

3. Le vocabulaire associé à "banque"

Ce vocabulaire est présenté en annexe 2. Au cœur de cet univers, figurent les noms des principales banques françaises puis internationales. Pourtant, le premier nom de pays apparaissant dans cette liste n'est pas *France* mais *Suisse* ! En fait, ce serait même *Grande Bretagne* si les articles n'hésitaient pas entre *Angleterre*, *Grande Bretagne* et *Royaume Uni*. Enfin, *New York* est la seule ville qui apparaisse dans cet univers : la géographie financière mondiale est sans équivoque !

Quelle est l'*activité* des banques ? Les *affaires* (généralement au pluriel) comme l'indique le substantif le plus significativement sur-employé. Ce mot étant fortement polysémique, la suite de l'univers de *banque* — comme les verbes : *détenir*, *diriger*, *accorder* (*crédit*, *prêt*), *regrouper*, *contrôler*... ou les substantifs : *filiale*, *investissement*, *gestion*, *fonds*, *compte*, *crédit*... — suggère quelques sens plus précis et un réseau sémantique cohérent qui sera décrit plus bas.

Ces listes restent assez abstraites, car les mots qui y figurent sont sortis de leur contexte. On propose donc un retour au texte, en demandant à l'ordinateur de sortir les passages les plus caractéristiques de cet univers. Pour cela, le programme relit l'ensemble du corpus et classe les phrases où figure le mot recherché, en fonction de la densité (absolue et relative) des associations y figurant (faute de place, l'annexe 1 donne seulement les premières phrases). Ces phrases peuvent être considérées comme les citations canoniques que donne tout bon dictionnaire à l'appui de ses définitions. Ici, le choix n'est pas fait arbitrairement par le lexicographe mais il est effectué objectivement, de telle sorte que ces phrases sont certainement les plus caractéristiques du mot recherché.

Deux remarques :

Une remarque de méthode. Dans le corpus entier, 11.1% des mots sont des verbes, dans l'univers de *banque*, cette proportion n'est que de 10,3 (-7,1%). En utilisant la formule (1), on va donc surestimer l'espérance mathématique des verbes dans l'univers de banque et quasiment tous les verbes auront des liens négatifs avec ce vocable… Pour éviter ce biais, il faut pondérer (corriger) l'espérance mathématique des verbes de -7,1%. On procède de la même manière pour toutes les autres catégories grammaticales.

La lecture de ces listes et de ces phrases suggère de nombreux sens spécifiques pour le mot banque. Aussi serait-il souhaitable de compléter l'analyse en procédant à la manière d'un dictionnaire, c'est-à-dire en regroupant de manière systématique les emplois attestés autour de quelques noyaux et en donnant, pour chacun d'eux, une définition et des exemples. Pour cela nous demandons à l'ordinateur de procéder comme les lexicographes : rechercher les synonymes et les phrases les plus caractéristiques de chacun des sens possibles.

B. Les synonymes de *banque*

Les corpus lemmatisés fournissent un outil efficace pour la recherche des synonymes d'un mot polysémique. Après avoir expliqué la méthode, les principaux synonymes seront présentés.

1. Taux de synonymie

On demande à l'ordinateur de procéder comme les lexicographes : rechercher les synonymes, les hyperonymes et les antonymes pour reconstituer les différents sens.

Rappelons que deux ou plusieurs mots différents sont dits synonymes lorsqu'ils partagent un ou plusieurs sens. La synonymie s'établit en substituant un mot à un autre dans le même contexte. Si le sens n'en est pas affecté, les deux mots sont considérés comme synonymes (voir ci-dessus, la notion de paradigme).

Rechercher les synonymes de *banque* consiste donc à se demander : quels substantifs peuvent lui être substitués, dans tout ou partie des phrases où il est employé, sans en changer le sens ? L'ordinateur ne peut répondre directement à cette question puisque la signification lui est inaccessible. En revanche, il peut dire quels sont, parmi les autres substantifs usuels du corpus, ceux qui partagent une partie significative de l'univers du mot pôle, en l'occurrence la *banque,* (associations positives et négatives). La démarche se déroule en trois temps.

1. On établit l'univers lexical du mot pôle (banque) qui vient d'être présenté ;

2. On demande de rechercher les phrases où ne figure pas le mot pôle (*banque*) mais au moins un autre substantif (synonyme potentiel) ;

3. Dans cet ensemble, on recherche les phrases où se trouvent le maximum d'associations positives et le minimum d'associations négatives caractéristiques, c'est-à-dire les passages où l'on est certain qu'il est question des banques même si ce mot n'est pas employé. On utilise pour cela un indice d'association ou "taux de synonymie" :

$$\text{Taux de synonymie} = \frac{\text{Nombre d'associations positives} - \text{Nombre d'associations négatives}}{\text{Taille de la phrase}}$$

Par exemple, en moyenne, les phrases contenant le vocable *banque* ont un indice moyen d'association de 0.21. Autrement dit, ce vocable "aimante" plus d'un mot sur cinq dans les phrases où il apparaît. Dans la recherche des synonymes, on ne retient que les phrases — **ne contenant pas *banque*** mais un autre substantif (synonyme potentiel) — dont le taux d'association (à l'univers lexical de *banque*) est supérieur à 0.21.

Secondairement, on relève aussi les adjectifs qui figurent dans des phrases semblables. En effet, en français, certains adjectifs peuvent être employés à la place d'un substantif.

2. Principaux synonymes de banque

Les résultats de l'expérience sont donnés dans le tableau ci-dessous (classés par taux de synonymie décroissante).

Les synonymes potentiels de *banque*

Synonymes potentiels :	Taux de synonymie
marché	0.334
groupe	0.316
activité	0.324
actionnaire	0.327
taux	0.347
établissement	0.401
président	0.301
opération	0.337
compte	0.329
intérêt	0.350
filiale	0.323
fonds	0.318
société	0.324
affaire	0.329
fusion	0.339
titre	0.331

Adjectifs	Taux de synonymie
français	0.328
grand	0.313
financier	0.335
public	0.334
européen	0.323
bancaire	0.348
important	0.355
monétaire	0.353
international	0.338

En deuxième colonne, la valeur de l'indice présenté ci-dessus, c'est-à-dire le solde relatif moyen des associations dans les phrases retenues (plus ce taux est élevé, plus la synonymie est probable). En dessous des substantifs, nous donnons, pour information, les adjectifs pour lesquels la même démarche a été effectuée.

Ce tableau appelle trois remarques :

— si la présence de mots comme *groupe*, *établissement*, *filiale* ou *société* est attendue (et prouve que la méthode est bonne), il peut paraître étonnant que *marché* soit présenté comme le meilleur synonyme de *banque*. La présence, dans la liste, d'autres mots comme *activité*, *taux*, *opération* ou *compte* laisse penser que l'expérience mélange les synonymes proprement dits avec les termes associés à l'univers de la finance. Faudrait-il alors inventer des filtres plus sévères ? Cependant, on se souviendra que la recherche ne porte pas sur la langue française mais sur le vocabulaire économique contemporain — c'est-à-dire sur des usages spécifiques. On peut penser que l'équivalence *banque(s)-marché(s)* est une synecdoque, figure rhétorique consistant à désigner les entreprises d'un même secteur économique par leur activité. Ainsi dira-t-on "l'automobile" pour l'ensemble des entreprises qui produisent,

commercialisent ou réparent ces véhicules. Dans le même ordre d'idée, il pourrait être logique que les observateurs de l'économie disent les "marchés financiers" pour désigner cet aspect du secteur bancaire… Dans ce cas, on en déduit que "marchés financiers" est un hyperonyme de *banque*.

— la courte liste des adjectifs suggère que la recherche des synonymes doit se faire également sur les groupes nominaux (du type "marché financier français" "grand établissement bancaire", "taux d'intérêt", etc.) Pour retrouver ces groupes, on utilise la méthode des syntagmes répétés (Pibarot & Labbé, 1998). Rappelons que le syntagme répété est une extension de la notion de "segment répété", extension que permet la lemmatisation du corpus. Par exemple, en utilisant cette approche, le meilleur synonyme de *banque* devient : "opérateur (ou) activité(s) (sur, de) (le, les) marché(s) financier(s)", ce qui confirme la synecdoque suggérée ci-dessus : dans certains cas, les acteurs de l'économie disent : *les marchés financiers* lorsqu'ils parlent des *banques*, vues sous l'angle de cette activité…

— la plupart de ces mots sont eux-mêmes fortement polysémiques. Il est donc probable que la synonymie n'est que partielle et qu'elle concerne seulement quelques emplois spécifiques (zone de synonymie) et que ces emplois se recouvrent plus ou moins ouvrant la voie à des regroupements. On aboutit ainsi à deux sens principaux dans les usages de *banque*. Le premier est dénommé à l'aide du syntagme le plus fréquent dans ce sous-univers : "opération(s) (de, sur les) marché(s) financier(s)".

C. Premier sens : *opérateur sur les marchés financiers*

Pour identifier les zones de synonymie et mesurer leur importance, on demande à l'ordinateur de relire l'ensemble du corpus pour reconstituer le vocabulaire caractéristique partagé par *banque* et par *marché*.

1. Trois situations

Le schéma ci-dessous récapitule les résultats obtenus avec *banque* et *marché*. La partie grise symbolise la zone de synonymie existant entre les deux mots. Cette zone est grossie pour les besoins de la représentation (en fait, elle ne représente que 15% de l'ensemble des emplois de "banque"). Elle correspond à trois situations différentes :

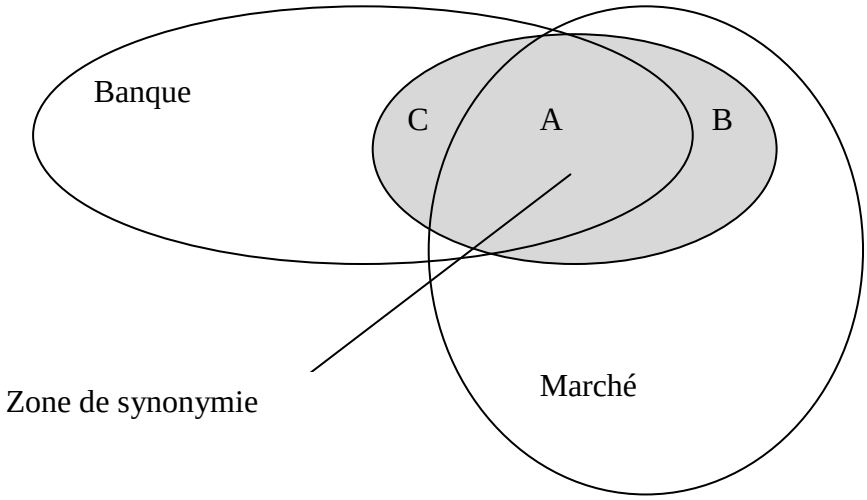
— A : 88 occurrences de *marché* (soit 3.5% du total de ses emplois), surviennent dans des phrases contenant également *banque* ;

— B : 169 des occurrences de *marché* (soit 6.5% des emplois du mot) apparaissent dans des phrases où ce vocable est employé dans des contextes semblables à ceux caractéristiques de la *banque* bien que ce mot soit absent ;

— C : 90 occurrences de *banque* (soit 7.3% des emplois de ce mot) figurent dans des contextes ne contenant pas *marché* mais qui sont caractéristiques de ce dernier.

Grâce au test statistique présenté ci-dessus, on détermine, au seuil de 1%, les vocables significativement sur-employés dans cette zone de synonymie.

La zone de synonymie entre *banque* et *marché*.



- Mots à majuscules : Paribas, Crédit Agricole, Société générale, Suisse, New York, Angleterre
- Verbes : contrôler, détenir, estimer, constater
- Substantifs : activité, opération, assurance, perte, portefeuille, spécialiste, taux, titre
- Adjectifs : anglo-saxon, britannique, domestique, financier, européen, français, immobilier, important, international, japonais, monétaire, étranger

Au total ce sont donc à peine plus de 10% des occurrences de *marché* qui peuvent être rattachées à l'univers sémantique de *banque* et 15% de celles de *banque* qui peuvent être considérées comme synonyme de *marché*… Certes, la faiblesse relative de ces taux s'explique en partie par le choix d'un seuil élevé qui conduit à ne retenir qu'un nombre assez restreint de phrases comme étant susceptibles de fournir des contextes communs. En contrepartie, on obtient des listes brèves et une information sans ambiguïté : les marchés sont d'abord *anglo-saxons*, voire *britanniques*, et le principal objet des banques ce sont des *activités* (*de marchés financiers*), des *opérations sur les marchés*...

En effet, l'analyse des syntagmes répétés dans la zone de synonymie indique qu'il s'agit, outre des "marchés financiers", des "opération(s) (de, sur les) marché(s) financier(s)", ou marché *international* ou *immobilier*, *domestique*, *des titres*, du *taux du marché*, etc.

2. Les antonymes de "opérateur sur les marchés financiers"

Pour établir un tableau complet, il faut aussi prendre en compte les vocables significativement sous-employés dans la zone grise du schéma (ils forment les antonymes de *banque* au sens d'*opérateur de marché*). La liste en est brève et éclairante :

- *Substantifs* : budget, cadre, entreprise, chaîne, chef, chiffre, chômage, emploi, formation, franc, heure, impôt, ingénieur, jour, magasin, mesure, mission, moyenne, nombre, patronat, production, salaire, salarié, site, syndicat, sécurité, temps, travail, usine.
- *Adjectifs* : humain, social.

Autrement dit, lorsque les observateurs parlent des activités de marché des banques, Ils "oublient" relativement des dimensions qui sont pourtant présentes dans leurs propos lorsqu'ils abordent d'autres thèmes. Au premier rang, figurent les relations professionnelles et sociales.

Enfin, tout bon dictionnaire comporte des citations illustrant le sens qui vient d'être défini. Le logiciel recherche les phrases les plus caractéristiques de l'intersection des deux univers (zone A du graphique ci-dessus). Voici la phrase qui contient à la fois *marché* et *banque* accompagnés du plus grand nombre de vocables associés positivement aux deux univers et le moins de vocables significativement sous-employés dans ces mêmes univers :

> *La complexité croissante des opérations financières, le développement de marchés sophistiqués et la concurrence qui pousse les banques à prendre toujours plus de risques incitent à s'interroger sur la capacité des banques centrales et des organismes internationaux à contrôler les systèmes bancaires.* (*Le Monde*, 6 février 1996).

Naturellement, un tel calcul "avantage" les phrases longues. On refait donc la même opération en rapportant le nombre d'associations à la longueur de la phrase. Les deux phrases suivantes sont celles qui obtiennent le meilleur score relatif :

> *Les banques d'affaires anglo-saxonnes s'imposent sur le marché français. (L'Expansion, 19 décembre 1996). Les activités de marché entraînent la banque Paribas dans le rouge.* (*Les Echos*, 29 février 1996).

Enfin, la présence de certains synonymes potentiels de *banque*, dans la zone grise du schéma permet de rattacher à ce premier sens, une série de substantifs du tableau 1 (et donc de paraphrases) plus spécifiques : *activité*, *taux*, *opération*, *titre*, *intérêt*… Pour chacune de ces significations particulières, on procède de la même façon : recherche de l'intersection entre les vocabulaires et recherche des phrases caractéristiques.

Avec ces éléments, il sera relativement aisé de rédiger le premier paragraphe de la définition du dictionnaire du vocabulaire de la presse économique concernant les banques comme *opérateurs sur les marchés financiers*. Le deuxième paragraphe concernera la banque comme *groupe financier*.

D. Autres sens du mot *banque*

En nombre de phrases concernées, le deuxième sens de banque est *groupe financier*. L'encadré ci-dessous récapitule le vocabulaire caractéristique de l'intersection entre *banque* et *groupe*, et présente les phrases les plus significatives de ce second sens.

Vocabulaire caractéristique des *banques* comme « groupes financier(s) »

Noms propres : BTP, CIC, Crédit Agricole, Crédit Lyonnais, Fiat, GAN, Italie, Natexis, New York, OPA, PME, Paribas, Rivaud,

Verbes : contrôler, devenir, diriger, recentrer, regrouper, détenir, estimer, financer, mener, peser, prendre, venir,

Substantifs : acquisition, actif, actionnaire, affaire, analyste, établissement, restructuration, assurance, assureur, banquier, bilan, compte, consortium, contrôle, crise, crédit, dette, difficulté, dirigeant, donnée, entité, exploitation, filiale, activité, finance, fonds, fusion, gestion, immobilier, intérêt, investissement, métier, opération, participation, partie, perte, portefeuille, privatisation, provision, rapprochement, redressement, rentabilité, risque, spécialiste, taux, titre,

Adjectifs : anglo-saxon, bancaire, belge, britannique, central, chinois, suisse, commercial, direct, domestique, européen, financier, français, grand, immobilier, important, international, italien, japonais, lourd, monétaire, privé, public, régional, étranger

Phrases caractéristiques :

« Une bonne raison à cela : aux Etats-Unis et en Grande-Bretagne (et au Japon aussi, mais les fonds japonais confient leurs fonds en gestion aux banques et aux compagnies d'assurance, avec lesquelles

elles ont parfois maille à partir), de très puissants fonds de pension détiennent une part considérable de la capitalisation boursière et ils ont, compte tenu des masses de fonds qu'ils gèrent, un pouvoir de discussion avec les entreprises. » (*Les Echos*, 30 janvier 1996).

« Plus pessimiste, il estime que la crise de l'industrie bancaire italienne va s'accentuer, en raison d'une compétition accrue entre les banques elles-mêmes et entre les banques et les nouveaux acteurs (compagnies d'assurances, grands magasins, banque directe), de la pression de plus en plus forte des mouvements de consommateurs, de la convergence des taux d'intérêt au niveau européen et de la stagnation de l'économie italienne. » (*Le Monde*, 29 avril 1996).

« Le partenaire idéal sera celui qui, comprenant que le groupe CIC apporte aux PME régionales, aux régions, aux métropoles françaises des services "plus" spécifiques, sans aucun doute exceptionnels, et qu'il a un fonctionnement tout à la fois efficace et décentralisé, dira : "ce groupe fonctionne, il a du potentiel, préservons son unité et ses banques régionales". » (*Les Echos*, 8 juillet 1996).

On procède de la même façon que ci-dessus : définition de la zone de synonymie — étendue et composition —, principaux syntagmes répétés puis recherche des phrases caractéristiques… A ce second sens, apparaît une série de synonymes plus spécifiques : *actionnaire*, *établissement*, *société*, *fusion*, *fonds*, *filiale*…

Avec ces deux grandes entrées, nous avons présenté l'essentiel du sens que donne au mot banque le vocabulaire économique contemporain. Toutefois, la liste des synonymes potentiels présentée dans le tableau 1 n'est pas totalement épuisée. Il reste quelques vocables qui ne peuvent être rattachés à l'un ou l'autre des deux principaux sens possibles que nous venons de décrire succinctement et qui fournissent des acceptions particulières que tout bon dictionnaire se doit de mentionner à la fin des principaux articles. Pour le mot *banque*, les plus intéressants sont incontestablement *président* et *affaire*. L'importance du premier s'explique par la synecdoque classique consistant à désigner une entreprise par son chef comme on le fait d'un pays par sa capitale. Mais ici, il ne s'agit pas du PDG, comme pour les entreprises non financières, mais du *président*, ce qui fournit un autre signe de la domination de la culture anglo-saxonne dans le domaine économique… Quant à *affaire*, la majorité des emplois de ce substantif ne correspondent pas au synonyme d'activité (on disait autrefois "banque d'affaires") mais à "scandale". Voici à tire d'exemple, la phrase la plus caractéristique de ce sens que les banques ne souhaitaient sans doute pas :

> *Ainsi, Barings a été la "victime" d'un opérateur de Singapour ; Daiwa a été "trompé" par le patron de sa filiale de New York ; le Crédit Lyonnais a perdu des milliards de francs dans le financement du cinéma américain par l'intermédiaire de sa filiale néerlandaise et les pertes les plus lourdes de la banque publique française proviennent, pour l'essentiel, de filiales mal ou pas contrôlées ; les déboires du Crédit Foncier sont aussi la conséquence des risques pris par des filiales*… (Le Monde, 6 février 1996)

Conclusions

Le travail qui vient d'être présenté est une œuvre collective. Les articles de journaux ont été rassemblés avec l'aide de J. Leselbaum, les programmes et les calculs ont été mis au point avec C. Labbé. De nombreux chercheurs nous ont remis des textes électroniques. Par exemple, les livres de Le Clézio ont été copiés par M. Kastberg-Sjöblom.

1. Sur la signification du mot banque. La signification du mot "banque" dans le vocabulaire de la presse économique contemporaine est maintenant établie de manière synthétique et précise. Deux conclusions principales se dégagent : d'une part, l'importance primordiale des marchés financiers et des modèles d'organisation anglo-saxons et, d'autre part, la dématérialisation des activités bancaires qui se traduit d'ailleurs par une quasi-disparition de la monnaie et des chiffres dans les articles de presse portant sur ces sujets. Ces conclusions sont éloignées des définitions données par les dictionnaires, mais il est vrai que l'expérience ne portait pas sur la langue générale mais sur le vocabulaire des journalistes spécialisés.

Ce corpus économique date des années 1996-1998 et les résultats ont été présentés pour la première fois en février 2002 : dix ans avant la catastrophe financière de 2008, les signaux d'alarme étaient allumés, les mécanismes du désastre étaient en place et les observateurs avaient conscience du dénouement probable. Cependant, l'analyse montre aussi une difficulté évidente à décrire l'activité bancaire, du fait de l'obscurité de plus en plus grande des mécanismes financiers. Cette difficulté peut expliquer l'aveuglement des autorités et leur incapacité à prévenir la crise.

2. *La statistique appliquée au langage* apporte de nombreux outils qu'on ne pouvait pas aborder dans le cadre restreint de cette communication. Par exemple, les calculs présentés lors de cette conférence peuvent aussi être appliqués à la comparaison de plusieurs auteurs et, par conséquent, à une question qui passionne les littéraires : l'attribution à un auteur connu de textes d'origine douteuse ou inconnue (Labbé 2004, 2009a et b).

3. *Sur la lexicographie assistée par ordinateur.*

L'ordinateur ne pourra jamais rédiger automatiquement un article de dictionnaire. L'exemple qui vient d'être présenté montre simplement que les grandes bibliothèques électroniques peuvent apporter une aide précieuse aux linguistes et aux lexicographes. L'outil pourrait également trouver des applications dans de nombreuses activités allant de la terminologie à la critique littéraire, en passant par la traduction assistée par ordinateur ou la recherche d'informations sur la toile…

Pour cela, il faudrait disposer d'un échantillon représentatif des usages du français contemporain, échantillon comparable au British National Corpus. Il n'existe rien de tel et les Français semblent y avoir renoncé…

4. Ce genre de recherche nécessite de *vastes collections de textes* couvrant une période de temps assez longue. Malgré son étendue, notre corpus n'atteint pas encore cette dimension critique. En effet, notre étude date d'il y a une dizaine d'années et l'on peut se demander quel est le sens du mot *banque* après 2008 ? J'aimerai vous répondre, mais c'est impossible. Il aurait fallu que notre bibliothèque électronique soit régulièrement actualisée. Les moyens et le temps nous ont manqué.

En effet, il faut que les corpus soient exploitables : les graphies doivent être normalisées puis rattachées à leur lemme et à leur catégorie grammaticale… Certes, ces opérations sont en grande partie réalisées par des automates. Mais une bibliothèque électronique vivante exigerait des moyens hors de notre portée... Nous espérons avoir suggéré combien cet investissement pourrait être rentable, comme diraient les financiers !

Enfin il faudra aussi que l'on admette enfin la principale proposition de Saussure : la langue est le trésor commun des usagers qui la parlent et non pas la chasse gardée des linguistes et des lexicographes.

Références

Nos travaux sont accessibles en ligne à partir de notre page personnelle ou sur le site "Archives ouvertes" du CNRS (HAL-SHS)

Benveniste Emile (1956). "La nature des pronoms". Reproduit dans Benveniste 1966, p.251-265.

Benveniste Emile (1958). "De la subjectivité dans le langage". Reproduit dans Benveniste 1966, p.258-266.

Benveniste Emile (1970). "L'appareil formel de l'énonciation". *Langages*, 17, p. 12-18.

Benveniste Emile (1966 & 1970). *Problèmes de linguistique générale*. Paris, Gallimard (rééd. 1980).

Blumenthal Peter & Hausmann Franz J. Eds (2006). "Collocations, corpus, dictionnaires". *Langue française*, 150, juin 2006.

Burnard Lou (1995). *Users Reference Guide for the British National Corpus*. Oxford : Oxford University Computing Services.

Cressot Marcel (1963). *Le style et ses techniques*. Paris : PUF (1ère édition : 1947).

Crowdy Steve (1993). "Spoken Corpus Design". *Literary and Linguistic Computing*, 8-4, p 259-266.

Fabre Cécile, Habert Benoît & Labbé Dominique (1997). "La polysémie dans la langue générale et les discours spécialisés". *Sémiotiques*. 13, décembre 1997, p. 15-30.

Gougenheim Georges, en collaboration avec Michea René, Rivenc Paul, Sauvageot Aurélien (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris : Didier. Réédition augmentée en 1964 sous le titre : *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

Gougenheim Georges (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier. Nouvelle édition revue et augmentée, Didier, Paris, 1977.

Guiraud Pierre (1950). *Les caractères statistiques du vocabulaire*. Paris : PUF.

Guiraud Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : PUF.

Hubert Pierre & Labbé Dominique (1995). "La structure du vocabulaire du général de Gaulle". Communication aux 3e journées internationales d'analyse des données textuelles. In Bolasco Sergio et al. *IIIe Giornate internazionali di Analisi Statistica dei Dati Testuali*. Rome : Centro d'Informazione e stampa Universitaria, 1995, tome II, p. 165-176.

Kastberg-Sjöblom Margareta (2006). *L'écriture de J.M.G. Le Clézio : des mots aux thèmes*. Paris : Champion, 2006.

Labbé Cyril & Labbé Dominique (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble : CERAT, décembre 1994 et juin 1997. Reproduit dans *Lexicometrica*, 3, 2001.

Labbé Cyril & Labbé Dominique (2005). "How to measure the meanings of words ? Amour in Corneille's work". *Language Resources Evaluation*. 39, p. 335-351.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahier du CERAT.

Labbé Dominique (2004). *Corneille et Molière. Table ronde 7e Journées d'Analyse des Données Textuelles*. Louvain-la-Neuve 11 mars 2004. Grenoble : CERAT-IEP.

Labbé Dominique (2009a). *Qui a écrit Tartuffe ?* Montréal : Monière-Wollank. Réédité sous le titre : *Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan*. Paris : Max Milo.

Labbé Dominique (2009b). *Qui a écrit Dom Juan ? Molière est-il l'auteur des pièces parues sous son nom ?* Communication au séminaire Mathématiques et Société, Université de Neuchâtel, 9 décembre 2009.

Leselbaum Jean & Labbé Dominique (2002). "Lexicographie assistée par ordinateur. Signification de *Banque* dans le vocabulaire économique". In Morin Annie et Sébillot Pascale (Eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*. Rennes : IRISA-INRIA, 2002, Vol. 2, p. 447-456.

Nelson Gerald (1997). "Standardizing Wordforms in a Spoken Corpus". *Literary and Linguistic Computing*, 12, 2 , p 79-85.

Pibarot André et Labbé Dominique (1998). "Les syntagmes répétés dans l'analyse des commentaires libres". in Mellet Sylvie (ed*). 4e Journées d'analyse des données textuelles*. Nice, 1998, p. 507-516.

Saussure Ferdinand de (1916). *Cours de linguistique générale*. Publié par Charles Bally et Albert Séchehaye avec la collaboration d'Albert Reidlinger. Réédition critique par Tullio de Mauro, Paris : Payot, 1993.

Savoy Jacques (2006). *"Les résultats de Google sont-ils biaisés ?"*, *Le Temps*, 9 février 2006 (article en ligne : http://members.unine.ch/jacques.savoy/Papers/PageRank.html).

Savoy Jacques (2010). "Discours électoral et discours présidentiel". In Bolasco Sergio et al. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, Vol 2, p. 827-838.

Silberztein Max (1993), *Dictionnaires électroniques et analyse automatique des textes : le système INTEX*, Paris, Masson.

Silberztein Max (1995). "Dictionnaires électroniques et comptage des mots". In Bolasco Sergio et al. *IIIe Giornate internazionali di analisi statistica dei dati testuali*, Rome, CISU, I, p 93-101.

Annexe 1

Densités des catégories grammaticales dans l'univers de banque (U) comparé au reste du corpus total (C-U).

Catégories	C-U (‰) (Corpus-Univers)	U (‰) Univers	U-(C-U)/(C-U) (%)
Verbes	110,6	102,7	-7,1
Formes fléchies	*64,2*	*58,0*	*-10,0*
Participes passés	*18,8*	*19,0*	*1,0*
Participes présents	*3,5*	*3,8*	*9,0*
Infinitifs	*24,2*	*21,9*	*-9,2*
Noms propres	47,4	50,7	7,1
Substantifs	220,0	228,5	3,9
Adjectifs	69,0	74,0	7,2
Adj, participe passé	*13,0*	*12,0*	*-8,1*
Pronoms	47,7	38,0	-20,3
Pronoms personnels	*24,2*	*17,5*	*-27,6*
Déterminants	234,9	235,9	0,5
Articles	*141,4*	*155,0*	*9,6*
Nombres	*68,2*	*56,8*	*-16,8*
Adjectifs possessifs	*12,0*	*12,8*	*6,1*
Adjectifs démonstratifs	*6,2*	*4,7*	*-24,4*
Adjectifs indéfinis	*6,9*	*6,7*	*-3,3*
Adverbes	48,2	43,8	-9,2
Prépositions	185,2	190,4	2,8
Conjonctions	35,0	34,3	-2,1
Conjonctions de coordination	*23,4*	*22,7*	*-3,2*
Conjonctions de subordination	*11,5*	*11,6*	*0,3*

Annexe 2

Univers lexical de « banque » dans le corpus « vocabulaire de la presse économique »

Nombre de vocables spécifiques à l'univers : 274
Nombre de spécificités positives : 180 soit : 3975 mots en surplus
Nombre de spécificités négatives : 94 soit : 1416 mots manquants

1. Vocabulaire significativement suremployé

(Seuil de 1%, classement par catégories grammaticales et indices de liaison décroissants).

Mots à majuscule : Paribas, Crédit Lyonnais, Rivaud, Crédit Agricole, Société Générale, Indosuez, CIC, BNP, Suisse, CDR, Italie, Lazard, Mediobanca, BTP, AFB, Natexis, Barings, JP Morgan, Crédit Mutuel, Comptoir des Entrepreneurs, Delmas, Lehman Brothers, Bâle, NatWest, Warburg, Pallas-Stern, Hokkaido Takushoku, Angleterre, Marsalet, BIP, Hongkong, Lucien, UBS, Douroux, BIANC, Goldman Sachs, GAN, Bolloré, Royaume-Uni, Crédit Foncier, OPA, Eurotunnel, Fiat, PME, New York,

Verbes : détenir, diriger, recentrer, demander, mener, accorder, appeler, avoir, regrouper, devenir, obliger, contrôler, imaginer, prendre, estimer, constater, peser, venir, financer,

Substantifs : affaire, activité, filiale, établissement, investissement, fonds, taux, compte, gestion, crédit, opération, assurance, risque, finance, perte, créance, caisse, prêt, contrôle, association, financement, agence, bilan, faillite, épargne, dette, immobilier, sauvetage, sicav, assureur, tutelle, détail, gouverneur, défaillance, commerçant, encours, défaisance, caution, abandon, rapprochement, analyste, consortium, recapitalisation, intérêt, entité, donnée, réserve, système, commerce, client, fusion, acquisition, mandat, proximité, institution, portefeuille, autorité, commission, président, provision, argent, privatisation, actif, dépôt, spécialiste, marché, ministère, département, dirigeant, actionnaire, banquier, métier, participation, rentabilité, redressement, titre, partie, union, difficulté, fédération, pouvoir, crise, dossier, organisation, exploitation, restructuration,

Adjectifs : grand, français, financier, central, public, bancaire, régional, international, important, commercial, britannique, vert, italien, immobilier, populaire, domestique, mutualiste, créancier, suisse, monétaire, étranger, anglo-saxon, direct, chinois, belge, privé, japonais, lourd, européen,

Pronoms : qui, que, celui, un,

Adverbes : ainsi, notamment, auprès, récemment, très, puis, enfin, tant, encore, trop, non, alors, longtemps, aussi,

Déterminants : le, leur, certain, deuxième**,**

Conjonctions et prépositions : de, et, par, comme, parmi,

2. Vocabulaire significativement sousemployé

(Seuil : 1%, classement par catégories grammaticales et indices de liaison décroissants)

Noms propres : Alain, Renault, Etats-Unis, Amérique, Américain, Français, PDG, France Télécom

Verbes : affirmer, produire, représenter, falloir, dépasser, investir

Substantifs : image, matière, qualité, économie, mission, industriel, cours, recette, recherche, temps, impôt, constructeur, milliard, géant, patronat, alliance, stratégie, chômage, technologie, ingénieur, salaire, croissance, firme, nombre, prix, entreprise, emploi, marque, usine, chiffre, production, virgule, jour, chaîne, salarié, succès, monde, budget, voiture, contrat, site, fabricant, magasin, jeune, travail, gamme, film, guerre, vente, moyenne, commande, formation, franc, sécurité, chef, syndicat, distributeur, lieu, consommateur, loi, cadre, dollar,

Adjectifs : nécessaire, beau, destiné, entier, jeune, seul, vendu, plein, unique, social, nouveau, humain, numérique,

Pronoms : tout, y, il, je, nous, vous,

Adverbes : pas, vite, bien, ne, moins,

Déterminants : tel, neuf, ce, cinq, mille, cent, deux, notre, dix, mon, quarante, cinquante, vingt, quatre, chaque, trois, même, six, huit, trente,

Conjonctions et prépositions : sur, contre, que, pendant, dès, à, chez, mais, en, près, pour, quand, voilà

3. Phrases les plus caractéristiques (en valeur absolue) :

Parmi les plus gros deals de l'année, outre la fusion record dans l'assurance, on retrouve le rapprochement entre le Crédit Local de France et le Crédit Communal de Belgique, l'acquisition de la banque Indosuez par le Crédit Agricole, et les OPA de Paribas sur la Navigation Mixte ou de la Caisse des Dépôts sur le Crédit Foncier (*L'Expansion*, 19 décembre 1996)

Le contrôle des banques devient un casse-tête pour les autorités de tutelle la multiplication des défaillances bancaires souligne la difficulté croissante, pour les banques centrales et les grands organismes internationaux, de mesurer les risques pris par les établissements financiers. (*Le Monde*, 6 février 1996)

Ainsi, Barings a été la " victime " d'un opérateur de Singapour ; Daiwa a été " trompé " par le patron de sa filiale de New York ; le Crédit Lyonnais a perdu des milliards de francs dans le financement du cinéma américain par l'intermédiaire de sa filiale néerlandaise et les pertes les plus lourdes de la banque publique française proviennent, pour l'essentiel, de filiales mal ou pas contrôlées ; les déboires du Crédit Foncier sont aussi la conséquence des risques pris par des filiales… (*Le Monde*, 6 février 1996)

Les nombreuses réformes du système de pensions faites par les équipes gouvernementales successives, le développement du travail indépendant, l'encouragement à l'accès à la propriété par l'intermédiaire de crédits hypothécaires dont l'octroi est lié à l'achat d'une police d'assurance-vie ont fait les vaches grasses des compagnies cotées en Bourse, des sociétés mutuelles ou des filiales des banques et des caisses de crédit hypothécaire (*Le Monde*, 24 décembre 1996)

Misant sur le plan de sauvetage, d'une valeur de 1.062 milliard de marks, qui passe par l'injection d'argent frais par les banques, des abandons de créances, une aide des collectivités locales, des ventes de terrains et un effort du personnel, la direction de KHD s'est désormais fixé pour objectif de recentrer l'entreprise sur son métier de base, les moteurs diesel, ce qui passe par un retour au nom qui a fait sa réputation, " Deutz ", lors de la prochaine assemblée générale (*Les Echos*, 22 août 1996).

4. Phrases les plus caractéristiques (en valeur relative) :

La banque de détail et de proximité restera une affaire nationale ", souligne monsieur Delmas Marsalet (*Le Monde*, 26 novembre 1997)

Les banques françaises ont aussi recentré leurs activités et simplifié leurs structures (*Le Monde*, 4 octobre 1997).

Encore récemment, la Banque Commerciale Privée (BCP) et surtout la banque Pallas-Stern ont ainsi déposé leur bilan. (*Les Echos*, 23 avril 1997)

La banque Rivaud avait créé un système de recyclage d'argent sale. (*Le Monde*, 26 juin 1997)

L'annonce de la fermeture de la banque Hokkaido Takushoku et surtout de la décision du gouvernement japonais de garantir les créances et les dépôts de l'établissement ont rassuré les investisseurs (*Le Monde*, 18 novembre 1997)

La banque verte poursuit d'ailleurs son plan de développement des activités de marché (*Les Echos*, 29 mars 1996)