Introduction à la résonance textuelle

André Salem

EA2290 SYLED – Université de la Sorbonne nouvelle – Paris 3 – France
salem@msh-paris.fr

Abstract

In order to study texts the structures of which are tightly linked such as : translations of the same text, questions and answers to them, learning material, simultaneous variations of text units through both parts of the corpus have to be considered. This paper presents a new area of textometric studies : *textual resonance* and provides an overview of the first applications of a new computing tool in various fields of textual studies.

Résumé

Lorsqu'on étudie simultanément des textes dont chacun entretient avec l'autre des rapports étroits (traductions d'un même texte, tours de paroles polémiques, données d'acquisition, etc.) il est utile de considérer les variations conjointes de différentes unités textuelles dans les deux volets du corpus. Cette étude présente les perspectives d'un nouveau domaine de la textométrie que nous appellerons la *résonance textuelle* et rend compte des premières applications à des domaines divers d'un outil informatique en cours d'achèvement.

Mots-clés : textes parallèles, statistique textuelle, textométrie.

1. Introduction

Dans le domaine des études textuelles, plusieurs situations de recherche nées dans des contextes parfois très éloignés conduisent à étudier simultanément des textes dont chacun entretient des rapports étroits avec l'autre du point de vue de sa structuration. Tel est le cas par exemple lorsqu'on considère un texte et sa traduction dans une ou plusieurs autres langues. C'est également le cas lorsqu'on considère un texte et ses différents états, versions ou brouillons, successifs ou simultanés dans une même langue ou encore les traces écrites d'une interaction verbale entre deux individus dans une situation didactique, dans une conversation ou un débat polémique.

L'étude des différents volets de ces ensembles de textes, souvent produits en simultanéité, s'enrichit de la prise en compte des relations nécessairement étroites qui lient entre elles certaines parties des deux textes. Suivant le principe qui veut que les outils informatiques doivent être pensés pour la plus grande généralité, nous proposons ici de considérer une situation très générale de cette mise en relation entre textes afin d'y adapter des outils d'exploration très généraux qui pourront ensuite être paramétrés pour des utilisations spécifiques dans des situations diverses[1].

[1] Les expériences dont nous rendons compte dans cette étude ont été réalisées à l'aide de plusieurs outils complémentaires qui appartiennent à la version actuellement disponible du logiciel Lexico3 diffusée par l'équipe SYLED. Les fonctionnalités spécifiquement dédiées à l'étude de la résonance textuelle devraient être disponibles dès la prochaine version.

2. Alignements bitextuels

Commençons par décrire une situation très générique dans laquelle deux ou plusieurs ensembles textuels sont mis en correspondance. Nous noterons ces ensembles T_1, T_2, … T_n. Nous supposons que chacun de ces ensembles est constitué par la suite des *occurrences* d'un ensemble fini de *types* (ou de formes) qui en constituent le *vocabulaire* V_i. À l'intérieur de chacun de ces ensembles, les occurrences d'items particuliers appelés *séparateurs* permettent de découper la suite des occurrences en fragments disjoints (par exemple des *phrases* et/ou des *paragraphes*). D'autres séparateurs appelées *jalons* permettent éventuellement de découper chacun des textes en un ensemble de *parties* qui peuvent correspondre à des découpages du texte que le chercheur a décidé de prendre en compte (chapitres, intertitres, tours de parole, etc.).

Dans le cas particulier de la mise en correspondance de deux textes dont l'un constitue une traduction de l'autre on appelle *alignement bitextuel* la mise en correspondance systématique de couples de fragments dont chaque élément appartient à l'un des deux textes mis en correspondance. On parle d'*alignement phrastique* lorsque les fragments mis en correspondance sont des phrases du texte, d'*alignement de paragraphes* si les unités alignées sont des paragraphes etc.[2]. On appellera *correspondances de zones* des ensembles de correspondances moins systématiques entre sous-ensembles appartenant à chacun des volets de l'ensemble bitextuel.

3. Induction et résonance textuelle

Le schéma de la résonance textuelle est illustré sur la figure 1. Cette figure représente un alignement réalisé entre deux volets d'un ensemble textuel dont on peut ignorer la nature dans un premier temps. Des fragments de texte appartenant à chacun des deux ensembles (ces fragments peuvent être des phrases, des paragraphes, des tours de parole, etc.) ont été mis en correspondance. À partir de cette mise en correspondance, toute sélection d'un sous-ensemble d'unités dans un des volets du corpus induit une sélection correspondante dans l'autre volet.

La sélection du sous-ensemble qui permet d'amorcer le processus de résonance peut être réalisée par différents moyens.

a) une sélection *topographique* constitue un sous-ensemble d'unités appartenant à l'un des deux volets exclusivement fondée sur leur localisation dans le texte (ex : les dix premiers paragraphes d'un texte qui en constituent l'introduction).
b) une sélection par *seuillage* se réalise à partir de considérations portant sur la quantité des occurrences d'un terme (ou d'une liste de termes) dans les fragments d'un des volets du bitexte (ex : ensemble des paragraphes qui contiennent des occurrences de formes commençant par le trigramme *soc*).

La sélection induite sur les fragments du second volet du corpus sert alors de point de départ à une deuxième étape qui verra la sélection des termes que leur présence (ou leur absence) rend caractéristiques pour cette dernière sélection.

[2] Les chercheurs qui travaillent dans le domaine de l'alignement automatique ont élaboré des procédures automatiques qui permettent d'aligner automatiquement les textes au niveau des phrases avec une relative efficacité (Cf. Veronis, 2000).

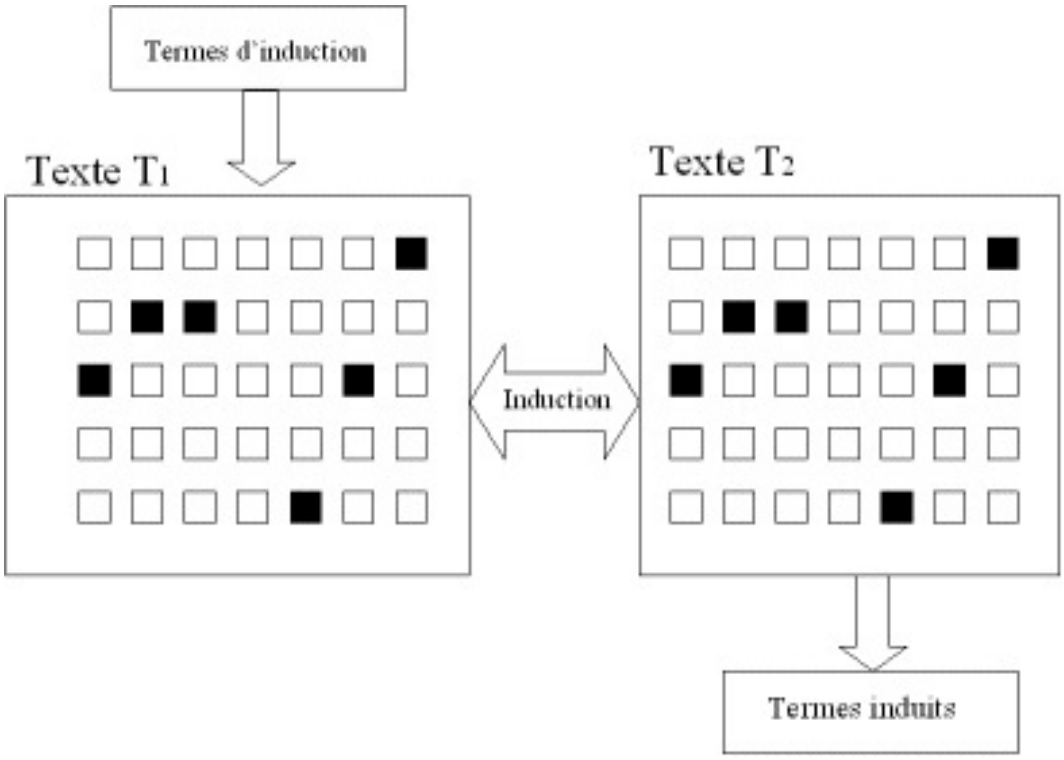

Figure 1.
Schéma général de la résonance textuelle
entre deux ensembles de textes

Guide de lecture pour la Figure 1 :

Deux ensembles textuels représentant, par exemple, deux traductions d'un même texte ont été soumis à un alignement par paragraphes. Dans chacun des volets T1 et T2 de la représentation ci-dessus chaque paragraphe du texte est représenté par un carré.

Un *terme d'induction* (par exemple un segment de texte correspondant à une locution dans le texte T1) a été présenté à un outil textométrique qui a distingué (en noir) les paragraphes du texte T1, dans lesquels ce terme trouve un nombre d'occurrences jugé important (par exemple : spécificité positive à un seuil fixé).

L'alignement en paragraphes, préalablement fourni, permet de sélectionner un sous-ensemble de paragraphes correspondants dans le second volet T2.

La dernière étape permet d'établir la liste des termes dont la fréquence est particulièrement remarquable dans le second ensemble sélectionné. Cette liste fournit des termes qui sont de bons candidats à l'appellation d'*équivalence de traduction*.

4. Applications à des domaines divers

Quelques recherches, dont certaines sont actuellement en cours, sont brièvement décrites dans ce qui suit. Ces recherches permettent d'illustrer la variété des applications possibles de la notion de résonance textuelle.

4.1. *Textes bilingues alignés*

On obtient de tels bitextes lorsqu'on met par exemple en correspondance un texte et sa traduction dans une autre langue[3]. La sélection dans l'un des volets du corpus des paragraphes dans lesquels les occurrences d'une forme, d'un segment répété ou de toute autre unité textuelle dépasse un seuil fixé provoque la sélection par *résonance* des paragraphes correspondant dans l'autre volet du corpus. La liste des unités textuelles particulièrement fréquentes dans cette seconde sélection met en évidence, dans l'immense majorité des cas des expressions qui sont des équivalences traductionnelles de la première expression.

Textes	A1 : texte juridique (en français)	A2 : sa traduction (en anglais)
Vocabulaires	V1 des formes du français	V2 des formes de l'anglais
Alignement	Alignement automatique au niveau des paragraphes entre A1 et A2	

Dans ce cas, par exemple, l'induction faite dans le volet français du corpus avec le segment répété *spéculer sur* (fr. 2 formes) renvoie le segment *speculate as to* (angl. 3 formes) qui en est la traduction. Ce type d'opération peut être réitéré en délimitant avec plus de précision la localisation des unités dont on recherche la traduction dans chacun des volets du corpus (Zimina 2000).

4.2. *Données d'acquisition*

Les textes B1 et B2 ont été recueillis dans une situation à caractère pédagogique. Un adulte converse avec un enfant à propos d'une lecture faite par ce dernier[4]. Trois séries d'entretiens ont été réalisées à plusieurs mois d'intervalle. Aux questions de l'adulte (B1) répondent les productions de l'enfant (B2). Le codage des tours de paroles permet de mettre en correspondance les questions de l'adulte et les réponses immédiatement fournies par l'enfant. Le but de la recherche est ici la mise en évidence d'interactions possibles entre le vocabulaire de l'adulte et celui de l'apprenant.

L. Sansonetti a étudié sur cette base la relation qu'on peut établir entre la fréquence des interrogations faites par l'adulte (*pourquoi ?,* etc.) et la variation ainsi que la structuration progressive dans le temps des réponses de l'enfant (*parce que*, etc.).

Textes	B1 : les productions de l'adulte	B2 : les productions de l'enfant
Vocabulaires	V1 le vocabulaire de l'adulte	V2 le vocabulaire de l'enfant
Alignement	Alignement réalisé au niveau des tours de parole de l'adulte et de l'enfant	
Partition	Une partition chronologique permet de suivre les progrès d'acquisition	

4.3. *Textes catégorisés*

À partir d'un texte C1, constitué cette fois par les réponses fournies à une question ouverte[5]

[3] Cet exemple est emprunté à des travaux menés par Maria Zimina au sein de l'équipe SYLED (Zimina, 2002). Pour une synthèse sur les problèmes du traitement des corpus parallèles, on se reportera à Veronis (2000).

[4] Cet exemple est emprunté à des travaux menés par Luiggi Sansonetti dans le domaine de l'acquisition au sein des équipes EA2290 SYLED et EA170 Calipso. (Cf. Sansonetti, dans ce même volume)

[5] Les données de cette enquête d'opinion dirigée par L. Lebart dans les années 70 ont servi de support à des exposés méthodologiques dans notre ouvrage collectif (Lebart et Salem, 1994). L'étude de l'apport des données de catégorisation a donné lieu à une publication avec B. Habert dans la revue *Traitement Automatique du Langage* (Habert et Salem, 1995).

par 2000 individus regroupés en cohorte d'âge et de diplômes équivalents, on a constitué un « texte » C2 qui contient la suite des catégories grammaticales, fournies par un outil informatisé appelé catégoriseur, qui correspondent à chacune des occurrences du texte B1. On réalise du même coup un alignement automatique qui concerne chacune des occurrences des deux textes, lesquelles se correspondent une à une.

Textes	C1 : texte des réponses à la question	C2 : catégorisation du texte C1
Vocabulaires	V1 des formes du français	V2 des catégories grammaticales
Alignement	L'alignement est alors réalisé au niveau de chaque occurrence du texte	
Partition	Les réponses sont regroupées en fonction de l'âge et du diplôme	

La sélection dans le second volet du corpus d'un patron syntaxique (ex : NOM + ADJ) permet de comprendre et d'illustrer par de nombreux exemples le fait que les personnes les plus diplômées produisent plus facilement des énoncés comportant des noms suivis d'adjectifs, les personnes moins diplômées privilégiant au contraire, la plupart du temps, les formes verbales.

4.4. *Débats télévisés*

En 1988, lors d'un face-à-face télévisé deux orateurs, F. Mitterrand et J. Chirac, prennent tour à tour la parole pour exposer contradictoirement leur point de vue[6]. On peut dans ce cas mettre en relation de résonance les tours de parole successifs (ou encore les tours de parole consécutifs, chacun étant relié à la fois à celui qui précède et à celui qui suit).

Cette mise en correspondance fait apparaître, entre autres choses, que l'emploi du mot *immigrés* par F. Mitterrand entraîne assez systématiquement celui du mot *immigration* par son contradicteur J. Chirac dans le tour de parole qui suit, et réciproquement.

F. Mitterrand :

il faut d' abord distinguer , c' est un problème qui a été vraiment exagéré et compliqué à plaisir. il y a plusieurs catégories de personnes visées par le débat actuel . il y a d - abord ceux qui ne sont pas des **immigrés** , qui sont les enfants d ' **immigrés** et qui sont nés sur notre sol . ceux - là ont vocation . ils sont français , sauf s ' ils en décident autrement à l' âge de dix - huit ans . il y a , ensuite , les naturalisés ; ce sont les **immigrés** qui désirent devenir français , là , l' administration étudie leur cas et il aboutit à reconnaître le droit à la naturalisation , selon son propre rythme . je n ' insiste pas. et puis il y a les **immigrés** _ ceux qui n ' ont pas envie de devenir français, qui veulent rester attachés à leur pays d ' origine _ de deux catégories : il y a les clandestins , et il y a ceux qui sont reconnus parce qu ' ils ont un contrat de travail et une carte de séjour . ceux qui sont clandestins , il n' y a qu ' une seule loi possible : il faut _ c ' est malheureux pour eux , mais c' est la nécessité _ il faut qu' ils rentrent chez eux et les dispositions doivent être prises , et elles ont été prises pour ceux - là , pour qu ' ils rentrent chez eux . et puis il y a ceux qui sont là avec leur contrat de travail et leur carte de séjour . est - ce qu ' il y en a trop ? ce que je sais , c' est que , dans les années qui ont précédé 1981 , il y a eu une formidable aspiration à faire venir chez nous des **immigrés** _ sans doute parce qu' on les payait moins bien que les autres , moins bien que les français , que les travailleurs français . /…/

[6] Une équipe du laboratoire de St Cloud animée à l'époque par M. Tournier a saisi ces débats télévisés afin de les soumettre à des analyses lexicométriques (Heiden et Tournier, 1998).

J. Chirac :

je voudrais répondre, moi , très clairement en m' appuyant sur mon bilan dans cette affaire ; parce que c' est très gentil de faire des promesses, mais enfin, encore faut il qu ' elles soient rendues crédibles par un bilan. s ' agissant de l ' **immigration** tout court , il faut la stopper , parce que nous n ' avons plus les moyens de donner du travail à des étrangers. aussi , naturellement , en supposant quelques souplesses naturellement, mais il faut la stopper . s ' agissant de l ' **immigration** clandestine, il faut évidemment lutter contre cette **immigration** avec beaucoup d ' énergie et reconduire les intéressés à la frontière ou les expulser. ils ont pris leurs risques en venant chez nous de façon illégale, ils sont le vivier naturel _ non pas en raison de leurs origines naturellement, mais parce que ce sont des marginaux et qui se cachent _ ils sont le vivier naturel des délinquants, voire des criminels : il faut donc les expulser. en 1981 - 82 - 83 , vous en avez régularisés 130000 : erreur capitale, car ça a été immédiatement un appel équivalent et même beaucoup plus large. nous , nous avons refoulé, en deux ans, plus de 130000 personnes, ce qui fait tout de même deux cents par jour, et je considère que ce n' est pas suffisant . nous le faisons , naturellement, en nous entourant de toutes les exigences de l' humanisme, de respect des droits de l' homme , mais c ' est une nécessité impérieuse. et puis nous devons nous protéger contre ces entrées. alors , je voudrais simplement poser une question. moi, j ' ai fait voter des lois pour la sécurité _ mais j' imagine que nous y viendrons tout à l' heure _ et contre l' **immigration** et notamment l' **immigration** clandestine, en particulier /…/

Tableau 1.
Exemple de résonance entre immigrés *et* immigration
lors du débat télévisé Mitterrand-Chirac
(présidentielles françaises 1988)

Comme on le comprend, ces deux vocables renvoient à des présentations différentes d'un même phénomène : le problème de personnes immigrées (i.e. déjà présentes sur le sol national) qu'il s'agit donc de gérer s'oppose à la présentation de *l'immigration* comme processus en cours et potentiellement menaçant.

Textes	D1 : les productions F. Mitterrand	D2 : les productions de J. Chirac
Vocabulaires	V on peut ici postuler l'existence d'un vocabulaire commun	
Alignement	Alignement réalisé au niveau des tours de parole des interlocuteurs	
Partition	Une partition chronologique permet de suivre la progression du débat	

La description de cette opposition bénéficie grandement du recours à des outils permettant l'étude directe de la résonance textuelle.

4.5. *Parentages de textes littéraires, recherche de plagiats*

L'étude des parentés/ressemblances/plagiats entre textes constitue un volet des études littéraires pour lequel les méthodes d'approche quantitative des textes sont souvent convoquées afin d'apporter un éclairage objectivant.

Textes	E1 : Un texte littéraire	E2 : Un second texte littéraire
Vocabulaires	V est le vocabulaire commun du français	
Alignement	Mises en correspondance expérimentales de zones textuelles par le chercheur	

Dans ce domaine, sauf contexte particulier, l'alignement entre différentes zones de textes dont on tente d'apprécier la proximité n'est pas donné mais posé à titre d'hypothèse sous la responsabilité du chercheur. La recherche de parentés entre les textes porte souvent sur l'agencement des unités lexicales mises en correspondance à l'intérieur des phrases. La résonance topographique constitue un outil privilégié pour le repérage de parentages.

5. Conclusion

Comme on le voit sur les quelques exemples qui précèdent, les applications du concept de *résonance textuelle* concernent un vaste ensemble de domaines de recherche liés aux approches formelles des corpus de textes.

Les applications du schéma général de la résonance textuelle à des problèmes de recherche particuliers permettent des comparaisons dont la nature peut varier d'un domaine à l'autre. Dans le cas de corpus dont l'alignement est suggéré par la nature des textes confrontés (traductions, versions différentes d'un même texte) la résonance permet de passer d'une liste de termes appartenant à l'un des volets à des termes qui lui correspondent dans le second à travers une transition fondée sur une correspondance topographique entre ces textes.

Dans le cas de confrontations verbales entre plusieurs locuteurs, la résonance peut permettre de juger de l'influence des productions de chacun des locuteurs sur celles l'autre. Dans d'autres cas (comparaison de textes littéraires etc.), l'appariement entre différentes zones appartenant aux différents volets du corpus constitue l'objet même de la recherche et doit être construit progressivement.

Les procédures informatiques qui permettent de réaliser ces expériences similaires dans des champs d'application très divers semblent par contre relever d'un modèle générique dont les paramètres pourront être modifiés par l'utilisateur en fonction de chaque projet de recherche.

Références

Habert B. et Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *Traitement automatique des langues*.

Heiden S. et Tournier M. (1998). Lexicométrie textuelle, sens et stratégie discursive http://lexico.ens-lsh.fr/biblio/slh/siad98/siad98.html

Lamalle C, Martinez W, Fleury S et Salem A. (2002). *Les dix premiers pas avec Lexico3. Outils lexicométrique*s. http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002.*

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

Veronis J. (2000). *Parallel text processing,* Kluwer Academic publisher.

Zimina M. (2002). Repérages lexicométriques des équivalences à basse fréquence dans les corpus bilingues. *Lexicometrica*, numéro spécial Corpus alignés (http://www.cavi.univ-paris3.fr/lexicometrica/)