# Modélisation d'un jeu de langage en vue d'explorations textuelles : l'exemple des chaînes de traitement de la plateforme *LinguaStream* appliquées au phénomène évaluatif.

Nous présentons dans cet article une grammaire locale de l'expression de l'évaluation, inspirée de la notion de jeu de langage de Wittgenstein. Le corpus étudié est un ensemble de critiques de livres déposées sur des sites commerciaux par des internautes. Une description linguistique des expressions récurrentes de ce discours est implémentée dans la plateforme de TAL *LinguaStream*. Cette implémentation, dont nous décrivons le processus, a pour fonction non pas de procéder à un traitement automatique des discours, mais de faciliter l'exploration des corpus textuels.

**Modelisation of a language-game in order to investigate corpora**

In this paper, we wish to consider the topic of evaluation in a corpus of book reviews written by readers on commercial websites. In particular, we study how evaluation may be expressed by regularly occurring sequences. We analyse some of the relevant linguistics patterns of evaluation in order to build a local grammar, influenced by the notion of *language-game* (Wittgenstein). This local grammar is formalized and implemented in a generic platform for Natural Language Processing, *LinguaStream*. The aim of the implementation is not to provide an automatic discourse analyzer, but an exploratory tool enabling linguists to investigate large corpora

Stéphane FERRARI, Université de Caen, Greyc CNRS UMR 6072, Modesco Stephane.Ferrari@info.unicaen.fr
Dominique LEGALLOIS, Université de Caen, Crisco EA 4255, Modesco, dominique.legallois@unicaen.fr

## Introduction

Cette étude est une illustration de la complémentarité entre une analyse linguistique d'un genre discursif et l'implémentation de ce traitement dans une plateforme informatique. Elle est donc le fruit d'une collaboration entre deux disciplines dont l'objectif commun est autant une réflexion théorique sur les niveaux de traitement pertinents, qu'un travail pratique d'analyse discursive et de constitution d'outils permettant l'exploration de larges données textuelles.

Le corpus de travail est constitué d'évaluations de livres déposées par des critiques non professionnels invités par des sites Internet commerciaux à exprimer leur avis à propos de leurs lectures. Il s'agit là d'une pratique discursive relativement nouvelle, encore peu décrite[1], qui fera ici l'objet d'une analyse linguistique partielle. Plus précisément, nous mettrons en évidence certains des aspects de ce discours dont la récurrence et la systématicité permettent de déterminer un *jeu de langage* particulier et de construire une *grammaire locale* de ce jeu. Ainsi, la phraséologie du corpus, couplée à des catégories conceptuelles (rôles des participants, types de valeurs) permet de décrire des constructions implémentées par la suite dans la plateforme LinguaStream – spécifiquement conçue pour le traitement informatique des textes. L'objectif est non seulement la réalisation de ce travail d'implémentation, mais aussi, l'exploration d'autres corpus à partir de chaînes de traitement préconstituées grâce à l'implémentation. S'il porte en effet avant tout sur des livres, notre travail doit pouvoir s'étendre par la suite, grâce à son dispositif, à d'autres objets (cinéma, jeux, musique, arts vivants), afin de pouvoir apprécier, sur des discours autres (par exemple, la critique professionnelle, ou le discours publicitaire), les constantes ou les variations des formes linguistiques employées mais aussi des types de valeurs convoquées. Nous donnons, dans la

[1] Voir cependant Legallois et Poudat (2008).

conclusion, l'exemple d'un travail en cours sur l'exploration d'un corpus de critiques datant du 19e siècle.

En premier lieu, nous présenterons le corpus travaillé, ainsi que le cadre théorique dans lequel s'inscrit ce travail. Dans un deuxième temps, nous donnerons des exemples de traitements linguistiques implémentables et présenterons alors la mise en œuvre qui a été réalisée à l'aide de la plate-forme de TAL (Traitement Automatique des Langues) LinguaStream. Nous détaillons les différents modules de la chaîne d'analyse proposée, dont la finalité n'est pas ici l'analyse automatique du phénomène, comme cela se fait habituellement en TAL, mais plutôt l'assistance du linguiste dans sa tâche d'observation du même phénomène sur de nouveaux corpus. Nous ouvrirons pour conclure la discussion sur les moyens proposés pour faciliter l'accès des linguistes à de tels outils complexes.

## 1. Corpus, jeu de langage et grammaire locale

### 1.1. Corpus

L'évaluation est un rapport fondamental à notre entour, que ce rapport soit verbalisé ou non, conscient ou non. Nous entendons par évaluation, l'affectation directe ou indirecte de valeurs, c'est-à-dire de sens, aux objets (y compris nous-mêmes), afin de les positionner, ou de se positionner dans un champ normatif. Le mode d'affectation est multiple : perception, jugement, estimation, critique, avis, mesure, point de vue, sentiment, etc. ; l'évaluation peut se manifester, sur une sorte de continuum vertigineux, par de « simples » réactions physiologiques (rougissement, nausée, etc.) ou, à l'autre extrémité, par un acte d'un haut degré institutionnel (le verdict d'un juge, la délivrance d'un diplôme, etc.). Il s'agit donc là d'un phénomène sémiotique majeur dont la portée est difficilement mesurable, tant le nombre d'objets auxquels il s'applique est étendu.

Il ne s'agit évidemment pas pour nous d'étudier l'évaluation en tant que telle mais, de façon beaucoup plus modeste, d'en identifier les principales manifestations dans une pratique circonscrite : l'avis porté sur les objets culturels, en l'occurrence des livres. Les avis analysés ici n'émanent pas de critiques « professionnels », mais de lecteurs / clients sur des sites Internet.

Le corpus de travail d'environ 100 000 mots est constitué de 440 critiques de livres (essentiellement des romans, mais aussi des essais) déposées sur les sites amazon.fr et fnac.com par les internautes. Exemple d'une critique :

*"J'irai cracher sur vos tombes" est un livre surprenant : le style est simple donc facile à lire mais le dénouement est insoutenable, voire vraiment immonde. Le combat de ce Noir "blanc" est louable et son envie de vengeance est aussi la notre mais il y a d'autres façons de se venger ou plutôt de décrire une vengeance... Un livre à ne pas mettre entre toutes les mains donc.*

Le corpus a été balisé sous format XML.

S'ils ne sont pas « académiques », les textes examinés ici n'échappent pas, loin s'en faut, à toute convention : ils s'inscrivent dans un genre, dans une pratique déterminée. Surtout, ils manifestent une constitution relativement stéréotypée qui favorise l'emploi d'unités récurrentes pré-données, préfabriquées et donc identifiables. Nous avons affaire à des routines discursives, à un jeu de langage, qui ne se caractérise pas par l'instanciation de règles, mais par des coups prévisibles (au moins pour certains d'entre eux).

### 1.2 Jeu de langage

Dans une perspective de réflexion sur ce genre émergent qu'est le discours évaluatif sur Internet, mais aussi dans une perspective plus pragmatique de modélisation informatique, la notion de jeu de langage proposée par Wittgenstein (1961, 1976) nous a paru fondamentale pour décrire les phénomènes inhérents au corpus et à son traitement. Le philosophe désigne par ce terme toute pratique sémiotique régulière : *commander*, *décrire un objet d'après ses aspects, prier, raconter une histoire,* etc. Ces pratiques sont des jeux dans le sens où elles possèdent leurs propres « grammaires » que les locuteurs ont apprises le plus souvent implicitement, dans leurs expériences discursives. Elles s'apparentent également à des genres, comme le remarque pertinemment Bouquet (1999). Ainsi, pour le discours évaluatif sur les livres, produit par des « critiques amateurs » sur des sites commerciaux, les prescriptions discursives restent implicites, et pourtant, chaque texte est parfaitement conforme au « genre ». Les jeux de langage sont des textes, c'est-à-dire des productions effectives constituées de *coups*. Nous qualifions de *coups*, les énoncés relativement préconstruits, repérables grâce à une (relative) stabilité lexico-grammaticale. Ce sont, à proprement parler, des actions – comme sont des actions de jeu des coups opérés par le joueur d'échec – c'est-à-dire des formes disponibles possédant une unité et une fonctionnalité. Enfin, la grammaire, au sens très précis de Wittgenstein, détermine les éléments en jeu dans les coups : pour notre

corpus, essentiellement les participants (évaluateur et évalué, dont les rôles peuvent se décliner de plusieurs façons – voir *infra*) et le champ axiologique – entendons par là les différents types de valeurs constitutives du jeu (par exemple, valeur éthique, valeur référentielle, etc. - voir *infra*).

Contrairement à la notion de ressemblance de famille, qui a eu en sciences du langage et plus largement en sciences cognitives, le succès que l'on sait, la notion de jeu de langage – parce qu'elle a subi la concurrence de celle d'acte de langage, a assez peu intéressé les linguistes. Elle possède cependant, selon nous, l'avantage sur la théorie d'Austin de mettre en évidence la pertinence des coups – que nous comprenons comme des unités non pas produites, mais reproduites, et la pertinence des genres. Surtout, elle permet de mieux décrire les énoncés authentiques dans une perspective fonctionnelle de la linguistique, comme, par exemple, la *grammaire de construction* ou la *grammaire des patterns* (Legallois et François (2006), Legallois (2007)). Enfin, mais cela dépasse de loin le thème de cet article, une linguistique des jeux de langage serait, selon nous, à même de montrer la nature émergente (et non *a priori*) de la grammaire[2].

### 1.3. Grammaire locale

D'une certaine façon, *grammaire locale* et *jeu de langage* sont pour nous des expressions synonymes. Proposée par Gross (1995), l'expression de *grammaire locale* a été reprise par les linguistes anglais Hunston et Sinclair (2000) à propos même de l'évaluation. La définition que propose Poibeau informe parfaitement la nature de ce type de grammaire :

*Une grammaire locale est une grammaire qui permet de décrire des liens entre éléments lexicaux ou syntagmatiques, généralement au sein d'une proposition donnée.(...) On désigne également par grammaire locale la description de l'ensemble des transformations qu'une unité peut subir tout en conservant intacte sa nature sémantique. Ces transformations sont syntaxiques (passivation, relativisation, pluralisation) mais aussi lexicales (remplacement d'un élément de l'unité par un synonyme, insertion de matériau étranger en position de modificateur d'un des éléments...)* (Poibeau, 2003 : 27).

La grammaire locale constitue une formalisation du jeu de langage. Elle est conçue comme un répertoire de séquences, entendons d'unités de l'ordre du mot, de la collocation, du syntagme, ou de la proposition, dont on détecte une certaine fréquence et un rôle important dans l'expression de l'évaluation. Ces séquences sont soit figées, soit semi-figées, ou apparaissent comme des constructions libres, mais leur fréquence montre qu'elles sont en fait

[2] Voir Hopper (1998).

« pré-fabriquées »[3]. L'identification de ces séquences a été opérée grâce à des outils « légers » comme *Lexico 3* (qui permet l'identification des segments répétés), et analysés par un concordancier (*Concapp*) afin de déterminer ou d'affiner des patterns (par exemple avec la fonction « joker » laissant apparaître des collocations discontinues). Évidemment, la disponibilité de ce stock d'expressions ne peut empêcher les variations, fort nombreuses, qu'il est nécessaire pourtant de neutraliser pour constituer des patterns généraux flexibles, applicables à d'autres textes. C'est donc sur le pari que le texte évaluatif est en partie empreint d'idiomaticité que nous pouvons concevoir un relevé et une formalisation des éléments pertinents.

L'identification des séquences – des coups – doit être couplée à une « grammaire conceptuelle » (au sens de Wittgenstein). Ce travail est relativement complexe, et peut être, à chaque reprise, perfectible. Par exemple, au mot *livre* ou au mot *roman* sont associés des rôles différents. En nous inspirant de Charaudeau (1993), il nous a fallu prendre en compte le fait que le livre est envisagé selon plusieurs facettes[4] :

***Le livre-genre*** : l'énonciateur évalue la conformité du livre à un genre (*ce n'est pas une autobiographie, mais un roman*) ;

***Le livre-histoire*** : l'énonciateur évalue la fable ;

***Le livre-message*** : l'énonciateur évalue la signification profonde du livre, son « message », son apport en termes d'idées, de réflexion, etc. ;

***Le livre-texte*** : l'énonciateur évalue la forme styliste et narrative, la composition du livre, son rythme.

De même, l'énonciateur se présente sous différents rôles :

***L'énonciateur-lecteur*** : l'énonciateur se représente dans son activité de lecture, en se constituant soit comme témoin archétype, soit comme témoin subjectif (*c'était bien la moindre des choses que je lise jusqu'au bout*).

***L'énonciateur-liseur*** : l'énonciateur évoque ici sa propre image de lecteur, ingrédient nécessaire et préliminaire à la lecture de l'œuvre évaluée (*Je persiste à lire Michel Houellebecq. J'avais lu "Rester Vivant" qui était plutôt dans un genre de poésie*).

***L'énonciateur-critique*** : l'énonciateur-critique construit sa légitimité à prescrire la lecture (*c'est LE meilleur livre romantique de la littérature française, n'en déplaise à Marc Lévy*)

---

[3] Cf. la notion de *prefabs* chez Erman et Warren (2000).

[4] Nous donnons ici, pour simplifier, les seuls rôles du livre et de l'énonciateur. Dans Legallois et Poudat (2008) la liste s'étend aux rôles discursifs de l'auteur et du destinataire.

Par ailleurs, les valeurs en jeu dans ce discours doivent être regroupées par types. Les typologies axiologiques de la lecture sont peu nombreuses. Nous nous sommes fondés (après les avoir évaluées[5]) sur les catégories distinguées par Dufays (2006)[6]. Rapidement :

***la valeur esthétique*** *(ou la beauté) concerne les qualités stylistiques et/ou rhétoriques du texte, ou si l'on préfère, sa poéticité, le travail de sa forme* ;

***la valeur référentielle*** *(ou la vérité) permet d'apprécier le réalisme du texte, sa conformité à ce que l'on considère comme la vérité* ;

***la valeur éthique*** *permet de se demander si le texte préconise d'une manière ou d'une autre des modèles de comportement conformes à l'idée qu'on se fait du bien moral ou, au contraire, s'il préconise leur transgression ;*

***la valeur signifiante*** *(ou la polysémie) permet de se demander si le texte est clair ou unifié ou au contraire riche, dense, complexe, multiple ;*

***la valeur informative*** *(ou la nouveauté) permet de se demander si le texte – sur le plan formel comme sur le plan du contenu – est innovant, original ou subversif, ou à tout le moins riche en informations ou à l'inverse, s'il est conforme à des connaissances ou à des canons familiers ;*

***la valeur psychoaffective*** *(ou l'émotion) permet de se demander si le texte est émouvant, s'il mobilise beaucoup d'affects et favorise par là la projection, voire l'identification du lecteur, ou au contraire s'il et neutre, impassible, distant.* (Dufays, 2000 : 282).

À cette liste nous ajoutons :

***La valeur-emprise***, c'est-à-dire le « contrôle » de l'objet sur le lecteur. L'évaluateur se dit « captivé » par le livre, saisi, envoûté. (*C'est une oeuvre envoûtante et captivante, superbement écrite (ah ! le style de Dumas...*)).

La cohérence de notre grammaire locale se mesure à la co-présence des trois paramètres : une séquence est couplée à un type de rôle et à un type de valeur.

## 2. Exemples

Précisons qu'il ne s'agit pas à proprement parler de constituer une grammaire exhaustive de l'évaluation ; plutôt, il s'agit d'établir, à partir des données extraites du corpus

[5] Dans Legallois et. Poudat (2008).

[6] D'autres typologies sont possibles, comme celle, plus générale, de Martin et White (2005) dans le cadre de la théorie de l'*Appraisal.*

de travail, puis formalisées et implémentées, des configurations qui, une fois projetées, ne couvriront certes pas toute la teneur évaluative d'un texte, mais consisteront des indices discursifs généralisés nécessaires à l'interprétation « humaine », et orientant des parcours interprétatifs. Rappelons enfin que l'objectif est d'illustrer la pertinence et les avantages d'une informatique outillée, et que par conséquent, le travail d'analyse linguistique présenté ici sera nécessairement partiel.

Pour les exemples suivants, nous donnons une notation qui reflète, non pas véritablement la composition syntaxique, mais « l'identité » fonctionnelle de la séquence.

### 2.1. Exemple 1

Les locuteurs expriment fréquemment l'idée que, saisis par la force du livre, *ils n'ont pu le lâcher*. Le cliché se réalise de différentes manières, comme le montrent les quelques lignes du concordancier :

```
t de la nuit. Pas question de lâcher le bouquin avant la fin. Et
écise.      Je n'ai pas pu le lâcher avant de l'avoir terminé. Je
is ouvert. On ne peut plus le lâcher, jusqu'à la fin. Un livre a
ère page, et on ne parvient à lâcher le roman qu'à la dernière page
```

Cette variation peut être neutralisée :

NEGATION - VERBE MODAL - lâcher - *évalué* - AVANT LIMITE

NEGATION *{ne…pas, pas question, hors de question, impossible, difficile, pas facile}*

VERBE MODAL *{pouvoir, réussir, parvenir}*

évalué *{le, le livre, le bouquin, le roman,* etc.*}*

AVANT LIMITE *{qu'à la dernière page, jusqu'à la fin, avant de l'avoir terminé, avant la fin}*

Cette complexité témoigne des différentes variations des réalisations. Cependant, nous encodons le seul noyau invariant. À chaque identification, un répertoire est disponible qui propose une visualisation des actualisations possibles (voir partie 3.1.2).

À cette expression est associée un rôle : le livre ou le roman est perçu ici comme *livre-histoire*. L'énonciateur, qui se donne soit comme un lecteur subjectif (emploi du *je*), soit comme un témoin archétype (emploi du *on*) se construit donc comme un énonciateur-lecteur. La valeur mobilisée est ici l'emprise.

### 2.2. Exemple 2

L'exploration du corpus permet de recenser des phrases évaluatives telles que :

*on a rarement aussi bien décrit les incidences morales et psychologiques de la misère.*

*Aucun livre de ma connaissance n'a jamais si bien démontré […] les dégâts […] que peuvent occasionner la vie* (sic !)

dont les variations lexico-grammaticales sont patentes : *rarement / jamais ; aussi bien / si bien ; décrire / démontrer,* etc.

Les rôles restent variables : *livre-message* ou *auteur-écrivain*. Deux types de valeurs se partagent la séquence : la valeur-référentielle, lorsque le verbe employé est un verbe de « représentation » (*dépeindre, peindre, rendre, décrire*, etc.), et la valeur-informative, lorsque le verbe est un verbe « argumentatif » (*démontrer, montrer*, etc.). Il nous faut donc tenir compte de ces différentes réalisations pour constituer la grammaire locale de la séquence.

Soit, donc, les notations suivantes :

**VALEUR-REFERENTIELLE** : [LIVRE-MESSAGE/AUTEUR-ECRIVAIN - *rarement / jamais – si / aussi – bien –* VERBES DE REPRESENTATION { *dépeindre, peindre, rendre, décrire,* etc.]

**VALEUR-INFORMATIVE** : [LIVRE-MESSAGE/AUTEUR-ECRIVAIN - *rarement / jamais – si / aussi – bien –* VERBES ARGUMENTATIFS { *démontrer, montrer,* etc}]

### 2.3. Exemple 3

Parmi les emplois de la suite *un livre qui* (soit en phrase averbale, soit introduit par le présentatif *c'est*), on repère des occurrences homogènes. Par ex. :

*C'est un livre qui fait avant tout réfléchir.*

*Un livre qui nous fait nous poser des questions...*

*Bref, c'est un livre qui invite à la réflexion*

La séquence, que nous formalisons ainsi

**VALEUR-INFORMATIVE** : [*un livre*$^{livre\text{-}message}$ *qui* - VERBES CAUSATIFS REFLEXION {faire réfléchir, donner à réfléchir, faire poser des questions, inviter à la réflexion, etc.}],

instancie la valeur-informative : il s'agit bien d'évaluer positivement l'effet du livre sur le lecteur, en termes d'apport intellectuel. *Un livre* est ici invariablement encodé comme livre-message.

Une séquence très proche actualise soit la même valeur (valeur-informative) :

*Claude Gueux nous donne énormément à réfléchir et rebondit sur moult autres débats*

**VALEUR-INFORMATIVE** : [LIVRE-MESSAGE/AUTEUR-ECRIVAIN - VERBES CAUSATIFS REFLEXION {faire réfléchir, donner à réfléchir, faire poser des questions, inviter à la réflexion, etc.}],

soit la valeur-éthique :

*Primo Levi nous donne une leçon de courage et de force vitale qui nous permettent de maintenir l'espoir sur la capacité de l'être humain pour survivre malgré la haine des bourreaux.*

**VALEUR-ETHIQUE** : [LIVRE-MESSAGE/AUTEUR-ECRIVAIN - VERBES ETHIQUES *{donner une leçon de courage, donner une leçon de civisme, etc.}],*

### 2.4. Exemple 4

*Encore un Boris Vian, qui, comme "l'Ecume des jours, bouleverse et emporte le lecteur dans un monde étrange.*
*l'auteur nous plonge dans une trame invraisemblable, nous gardant en apnée tout en nous faisant de temps en temps remonter à la surface…*
*Une épopée familiale qui nous transporte dans d'autres contrées, dans d'autres époques*

**VALEUR-INFORMATIVE + VALEUR-EMPRISE** : [LIVRE-HISTOIRE / AUTEUR-ECRIVAIN – *nous*$^{ENONCIATEUR\text{-}LECTEUR}$ - *VERBES TRANSPORTS* {*emmener, transporter, entraîner, embarquer, plonger*} - *dans* – NOMS UNIVERS {*un monde, un univers, une histoire, des contrées, une atmosphère,* etc.}.

Cette construction trivalencielle locative, qui met en jeu un petit nombre de verbes {*emmener, transporter, entraîner, embarquer, plonger*}, un locatif sémantiquement déterminé mais variable {*monde, univers, ambiance, histoire,* etc.}, évalué par un adjectif, est un « coup » employé dans le seul jeu de langage qu'est le discours évaluatif. Les rôles sont stables : livre-histoire, auteur-écrivain, énonciateur-lecteur, et les valeurs mobilisées constituent la valeur-informative - elle caractérise l'habileté de l'écrivain, non pas à dépeindre fidèlement une réalité (qui serait donc reconstituée), mais à construire un monde personnel, inédit – et la valeur-emprise, car les verbes connotent le magnétisme de l'œuvre, qui permet au lecteur d'oublier son quotidien pour s'abandonner à un nouvel univers. Il s'agit donc là de

l'attestation d'une lecture *participative*, qui s'oppose à une lecture *distante*, selon la classification des analystes littéraires.

Cette même valeur-emprise se construit dans une construction proche :

**VALEUR-EMPRISE** : [ *on*$^{ENONCIATEUR-LECTEUR}$ *– se laisser -VERBES EMPRISES* { *entraîner, embarquer, bercer, prendre, surprendre, emporter*}].

dont voici les lignes du concordancier :

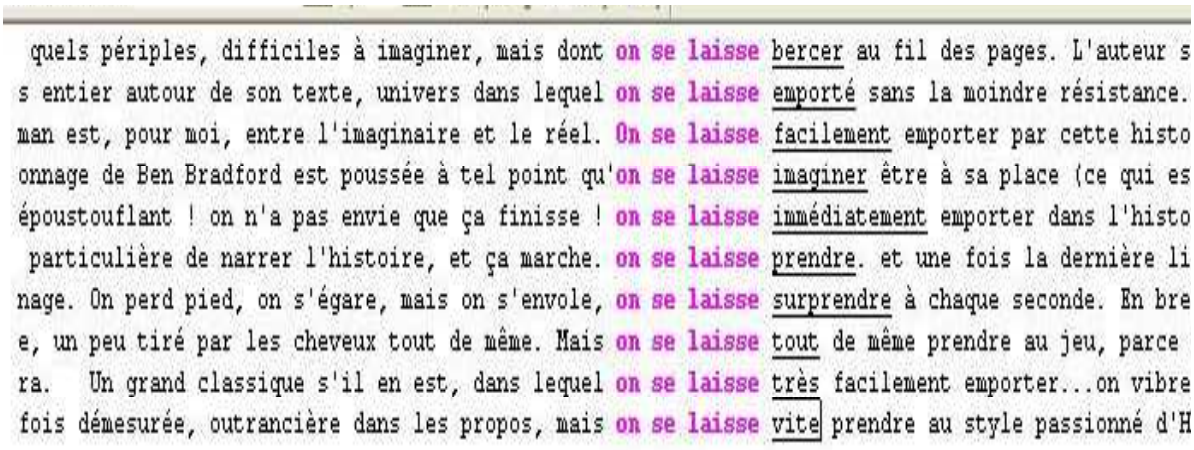
quels périples, difficiles à imaginer, mais dont on se laisse bercer au fil des pages. L'auteur s
s entier autour de son texte, univers dans lequel on se laisse emporté sans la moindre résistance.
man est, pour moi, entre l'imaginaire et le réel. On se laisse facilement emporter par cette histo
onnage de Ben Bradford est poussée à tel point qu'on se laisse imaginer être à sa place (ce qui es
époustouflant ! on n'a pas envie que ça finisse ! on se laisse immédiatement emporter dans l'histo
particulière de narrer l'histoire, et ça marche. on se laisse prendre. et une fois la dernière li
nage. On perd pied, on s'égare, mais on s'envole, on se laisse surprendre à chaque seconde. En bre
e, un peu tiré par les cheveux tout de même. Mais on se laisse tout de même prendre au jeu, parce
ra. Un grand classique s'il en est, dans lequel on se laisse très facilement emporter...on vibre
fois démesurée, outrancière dans les propos, mais on se laisse vite prendre au style passionné d'H

**Figure 1 : concondances de *on se laisse +inf.***

Ces quelques exemples, pris parmi une multitude, composent donc le répertoire de la grammaire locale de l'évaluation. L'enjeu, on l'aura compris, est de neutraliser les différences de réalisations, syntaxiques et lexicales, de ces séquences et de leur assigner les rôles conceptuels ainsi que les types de valeurs qu'elles expriment.

## 3. Mise en œuvre à finalité observatoire

Les premières observations sur corpus permettent déjà de dessiner les grandes lignes d'un modèle informatique pour une mise en œuvre de ce que l'on pourrait nommer un outil d'aide à l'observation. Nous décrivons ici une expérimentation qui, s'appuyant sur les régularités observées, permet leur repérage en corpus de manière automatique, ainsi que leur annotation.

Insistons sur le fait que ces repérages et annotations ne visent pas, à ce stade, une analyse automatique du phénomène, mais ont au contraire pour objectif d'outiller le linguiste d'un nouveau logiciel d'observation pour des études ultérieures sur de nouvelles données textuelles. Nous renvoyons le lecteur à Vernier, Ferrari et Legallois, 2007 et Vernier et al. 2007, pour une description d'autres mises en œuvre du modèle, à visée cette fois purement applicative dans le domaine de la fouille d'opinion.

### 3.1. Chaîne de traitements pour observer l'expression de l'évaluation

Afin d'expérimenter le modèle sur corpus, nous avons opté pour l'utilisation de LinguaStream, une plate-forme de TAL développée au GREYC (Université de Caen) qui permet notamment l'utilisation dans une même chaîne de traitements de différents formalismes[7]. L'objectif est ici de réaliser un outil informatique facilitant l'observation des régularités lexico-grammaticales précédentes, tant sur le corpus d'étude original que sur de nouvelles données, et permettant un enrichissement incrémental des règles ou lexiques sur lesquels reposent les analyses.

Une expérimentation comme celle que nous proposons implique de reformuler l'ensemble de nos observations précédentes, à caractère plutôt descriptif, en un modèle opératoire, à caractère prescriptif, comme montré dans Ferrari *et al.* (2005). Les formalismes mis à disposition dans LinguaStream laissent une grande liberté dans l'expression du modèle opératoire, qui peut être mis en œuvre tant à l'aide d'automates de type expressions régulières que de grammaires de type Prolog. Nous avons tiré parti de cette offre, certains types d'analyse étant mieux adaptés à la mise en œuvre des patrons lexico-grammaticaux, d'autres à la « remontée » d'informations sémantiques depuis un lexique jusqu'à des éléments textuels.

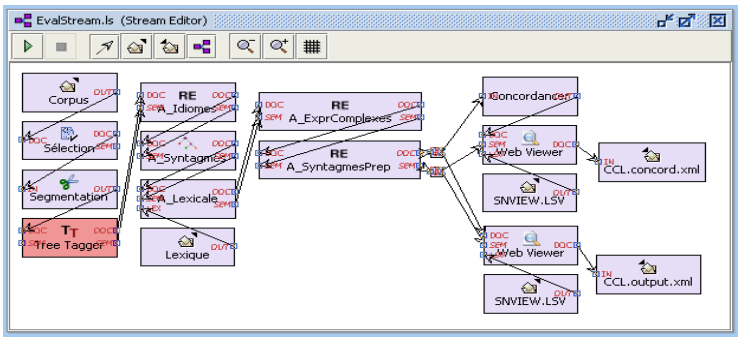


**Figure 2 : chaîne de traitements LinguaStream**

[7] Cf. Widlöcher et Bilhaut, (2005) et Enjalbert (2005).

### 3.1.1. Prétraitements

La chaîne LinguaStream de la figure 2 montre les différents composants utilisés pour cette expérimentation. Chaque boîte y représente un composant ou une ressource, les flèches entre les boîtes représentent la transmission d'information entre composants.

La première colonne de composants consiste en une série de « prétraitements » préparant aux analyses suivantes :choix des données en entrée, le « Corpus » au format XML, « Sélection » des éléments XML pertinents de cette ressource pour les analyses ultérieures (dans notre cas, nous concentrons les analyses sur le titre et le corps des avis, les informations concernant par exemple les dates et les auteurs des avis seront ignorées par les analyses menées ultérieurement), segmentation en mots («Segmentation») et catégorisation grammaticale à l'aide de *Tree Tagger* (Schmid, 1994).

À l'issue de cette première colonne de composants, la chaîne d'analyses se poursuit avec la transmission de deux informations en parallèles : une version du document d'origine enrichi au fur et à mesure d'ancres permettant d'y repérer les différents éléments analysés, et les résultats des analyses, transmis en parallèle et codés dans un fichier indépendant lors d'une sauvegarde. Cette première colonne de composants influence la qualité des résultats des composants dédiés à la mise en œuvre de notre modèle, dans la mesure où ils exploitent une partie des informations qui y ont été produites.

### 3.1.2. Repérage de formes

La deuxième colonne contient un ensemble de composants destinés au repérage des formes récurrentes observées. La boîte « A_Idiomes » exploite des automates, le symbole RE indiquant, dans la plate-forme, l'utilisation d'un composant de type expressions régulières ou « macro expressions régulières ». Les macro expressions régulières permettent de travailler sur des unités précédemment analysées en combinaison avec le formalisme des expressions régulières classiques. Nous renvoyons à (Widlöcher et Bilhaut, 2005) pour plus d'informations sur ces points. Ce composant effectue ici une amorce de l'analyse des formes lexico-grammaticales, en s'appuyant sur la présence de certains mots dans un certain ordre, avec éventuellement vérification de la catégorie grammaticale telle qu'issue du Tree Tagger si une ambiguïté semble devoir être levée. Par exemple, la structure récurrente *« impossible de lâcher »* s'y traduit par la règle déclarative suivante :

```
<idiom> impossiblede() %[0-1] {lemma:"lâcher"} </idiom>
```

```
/sem {synt:SNObj, sem:emprise, eval:idiom}
```

et une règle intitulée « impossiblede », exploitée par la précédente, disponible pour d'autres :

```
("impossible" "de"| "pas" "question" "de"| "difficile" "de"|
{lemma:pouvoir} %[0-1]|{lemma:parvenir} %[0-1] "à")
```

La première règle permet de marquer comme élément *idiom* un mot dont le lemme est « lâcher » et une expression le précédant qui est repérée par la règle intitulée « impossiblede », avec un mot supplémentaire pouvant s'intercaler (`%[0-1]`[8]). L'information qui est associée à l'élément découvert suit le mot-clé `/sem`. C'est une structure de trait renseignant sur la nature de l'élément repéré et/ou précisant quelle analyse mener ensuite pour compléter le patron : *eval:idiom* permet de caractériser ici un type de résultat de l'analyse de l'évaluation, *sem:emprise* précise la valeur à associer à l'expression repérée, *synt:SNObj* sera utilisé par un composant ultérieur pour associer un syntagme suivant l'expression repérée. La deuxième règle permet une variabilité lexicale sur une partie du patron. Les deux dernières entrées s'appuient directement sur le lemme (« parvenir », « pouvoir »), information issue du Tree Tagger, afin d'éviter une série de formes fléchies.

La deuxième boîte, « A_Syntagmes », représente un composant d'analyse de syntagmes nominaux. Il s'agit d'une grammaire Prolog dans laquelle nous avons injecté une partie de l'information lexicale liée à notre modèle[9]. Plus précisément, les clauses exploitent le formalisme GULP, proposé par (Covington, 1994), pour permettre la manipulation en Prolog des structures de traits. Les analyses menées dans ce composant s'inspirent des analyses de type « chunking », et fournissent aux composants ultérieurs des entités annotées à la fois d'informations « syntaxiques » (les catégories des mots au sein du chunk, sans le nombre et le genre) et sémantiques (celles issues du lexique préconstruit par nos soins et reflétant les observations précédentes). Ce composant s'appuie essentiellement sur les résultats du tagger pour composer les chunks, sans prendre en considération le genre et le nombre mais uniquement la catégorie grammaticale. Ce parti pris permet de traiter le texte tout-venant dans lequel nous avons remarqué de nombreuses fautes d'accord au sein des syntagmes nominaux, mais il souffre par ailleurs des faiblesses de la catégorisation grammaticale du Tree Tagger. C'est un compromis satisfaisant pour l'objectif de cette mise en œuvre, mais il serait nécessaire d'envisager un autre type d'analyse pour la mise en place d'un outil de traitement automatique.

---

[8] Le symbole % concerne plus précisément un élément considéré comme grain d'analyse pour le composant en cours de LinguaStream, qui est fixé pour l'analyse présentée ici aux mots tels qu'issus du processus de segmentation.

[9] Ce composant est réalisé en collaboration avec T. Charnois, GREYC

Les deux dernières boîtes de la deuxième colonne, « A_Lexicale » et « Lexique », représentent un complément d'analyse lexicale permettant de compléter l'information précédente notamment pour la catégorie verbale, qui n'est pas actuellement exploitée par le module d'analyse des syntagmes.

### 3.1.3. Complétion des séquences

À cette phase de l'analyse, toute information lexico-sémantique susceptible de concerner l'expression de l'évaluation est exploitée. Les éléments de séquence s'appuyant sur un lexique stable ont été repérés et annotés avec deux types d'informations : celles relatives aux valeurs associées, d'autres éventuelles pour guider les analyses restant à mener. La plupart des séquences sont en effet à compléter, soit par le rattachement d'un syntagme prépositionnel à un élément préalablement repéré, soit par regroupement d'éléments déjà annotés, etc.

Le traitement de tous les résultats précédents se fait par les deux composants de la troisième colonne. Le premier, « A_ExprComplexes », consiste en une analyse des expressions encore incomplètes. Un rattachement s'effectue à l'aide de macro expressions régulières, permettant de tenir compte de la discontinuité inhérente à certaines des structures observées. Lorsque le rattachement n'aboutit pas, c'est-à-dire qu'aucun syntagme du type attendu n'est trouvé dans la phrase, l'expression initialement repérée est actuellement abandonnée.

Le second composant effectue le rattachement prépositionnel au sein des syntagmes nominaux. Cette analyse complète le *chunking* précédent. Pour des raisons d'efficacité, elle n'est mise en œuvre que pour les syntagmes effectivement repérés comme impliqués dans une séquence évaluative.

### 3.1.4. Affichages pour l'observation des séquences repérées

La suite des composants est quant à elle d'une nature toute différente. Il s'agit de filtrages et de préparations à l'affichage. Les informations précédemment associées aux syntagmes ou chunks sont ici oubliées, à l'exception de celles directement en rapport avec notre étude, pour un affichage des résultats dans un navigateur – les informations issues du tagging et du chunking alourdissent inutilement les fichiers dans ce genre de tâche. Nous ne détaillerons pas les modules utilisés ici, qui reflètent des spécificités de la plate-forme LinguaStream que nous nous contentons d'utiliser : surlignage des éléments annotés, préparation à une sortie de type concordancier, mise en place de divisions cachées au format

HTML pour afficher dans un navigateur le corpus et, sur demande, les structures de traits associées aux éléments repérés, etc. Nous présentons des exemples de sortie pour illustrer ces aspects relatifs aux résultats.

### 3.2. Exemples de résultats

Afin de mettre en œuvre les analyses présentées plus haut, le corpus d'origine a été préalablement transcodé en XML, selon les méthodes préconisées par Habert et al. (1998). Il contient désormais des informations sur les éléments logiques des avis, selon leur disponibilité : *titre, date, lecteur diffusant l'avis, titre et auteur du livre visé*… La plate-forme LinguaStream permet de conserver ces annotations initiales lors d'une chaîne de traitements, celles issues des analyses étant insérées à l'aide d'éléments nommés[10]. Une feuille de style de type CSS qui aura été appliquée au corpus initial pour en rendre la lecture plus aisée peut donc encore être utilisée lors des observations, les indications de mise en forme demeurant valides pour la version analysée du corpus.

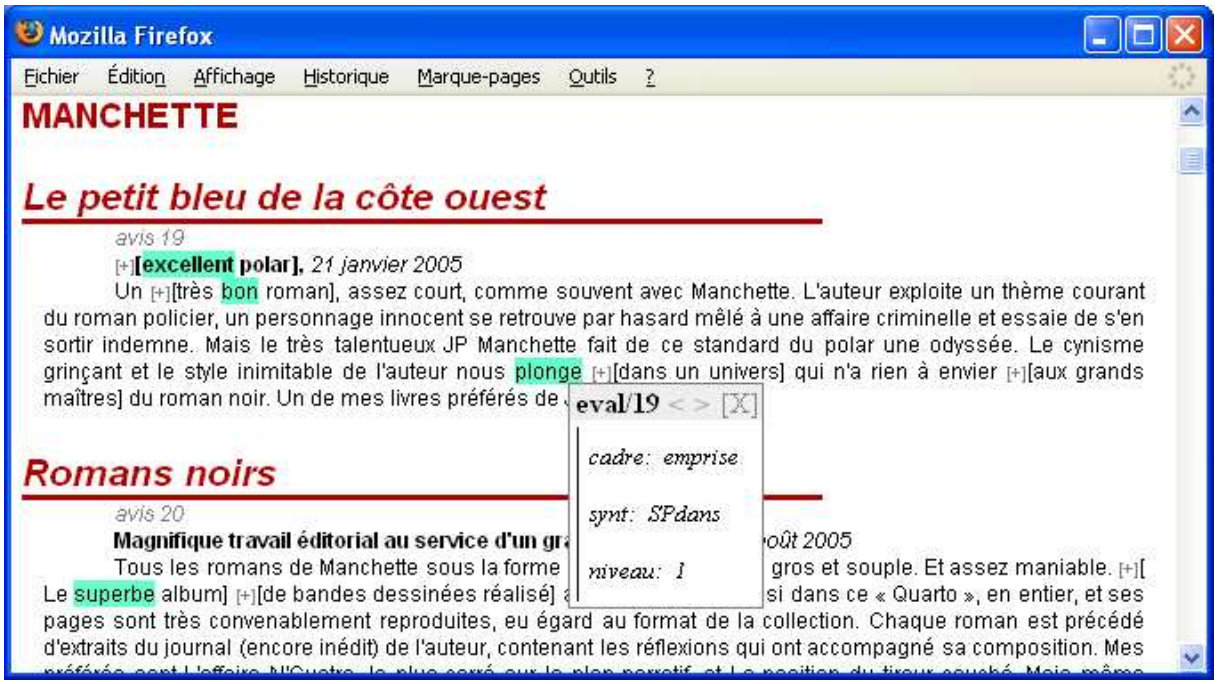


**Figure 3 : Exemple de résultat – un syntagme annoté**

La figure 3 permet d'apprécier le premier type d'affichage des résultats que nous obtenons avec la chaîne d'analyse précédente. Il s'agit ici du corpus original enrichi

[10] Notion d'élément nommé de la grammaire XML, consistant ici en un préfixe « ls » qui renvoie à l'*espace de nom* (namespace XML) relatif aux analyses issues de la plate-forme LinguaStream.

d'annotations. Des expressions y apparaissent en surbrillance, certaines apparaissent entre crochets et précédées d'un symbole [+]. Les surbrillances correspondent aux termes des séquences repérées qui sont directement issus des lexiques utilisés pour l'analyse. Les crochets entourent quant à eux des expressions ayant fait l'objet d'une analyse locale approfondie, car reconnues comme appartenant à une séquence.

Dans la copie d'écran proposée, le verbe *plonger* apparaît avec un surlignage reflétant son appartenance au lexique de l'évaluation. Le syntagme prépositionnel qui le suit apparaît encadré et précédé d'un symbole [+]. Il s'agit d'un syntagme traité par le module d'analyse des expressions incomplètes, la structure de traits associée au verbe plonger, *{cadre:emprise, synt:SPdans, niveau:1}*, précisant un rattachement d'un syntagme introduit par *dans*.

Le symbole [+] correspond, dans les corpus annotés par LinguaStream, à une division cachée contenant la structure de trait associée à l'élément qui suit ce symbole. Il est possible de cliquer sur ces symboles et sur les éléments en surbrillance pour ouvrir un cadre tel celui de la figure (eval/19) pour afficher des informations sur les éléments annotés. L'ensemble des résultats est donc consultable de manière interactive : un premier regard pour repérer les éléments annotés ; une analyse plus précise des structures de traits à la demande.


**Figure 4 : Affichage des résultats sous forme de concordancier**

La figure 4 permet quant à elle d'apprécier le deuxième type d'affichage possible prévu dans la chaîne d'analyse proposée. Il s'agit d'une vue partielle des résultats, inspirée des concordanciers classiques, qui permet de sélectionner un type d'annotation particulière et

d'afficher toutes les séquences de ce type trouvées dans le corpus. L'exemple de la figure a été construit sur l'élément de structure de trait associé aux annotations : *eval:idiom*. Il s'agit ici d'une catégorie d'expressions de notre grammaire que nous avons regroupées pour des similitudes dans l'analyse qui en est faite.

Le linguiste utilisateur de la chaîne dispose en réalité d'une grande liberté dans le réglage des paramètres de ce pseudo-concordancier afin de faciliter ses observations. N'importe quelle partie d'une annotation peut en effet servir au filtrage des expressions à afficher.

Il serait vain de nier la complexité du traitement tel qu'il a été exposé ici. Mais cette complexité est à la fois le reflet indirect de la labilité des formes constituant la grammaire locale, et le prix à payer pour un traitement satisfaisant de cette grammaire. Ainsi, le traitement modulaire informatique peut formaliser en partie le caractère phraséologique et holistes des formes linguistiques.

## Conclusion

Notre objectif initial était, dans le cadre d'une collaboration entre linguiste travaillant sur corpus et informaticien spécialiste du TAL, de parvenir à proposer un outil qui facilite les observations ultérieures du linguiste, tant sur le corpus initial que sur de nouvelles données. Cette facilitation des observations est directement corrélée avec la complexité de l'implémentation.

Il convient ici de modérer l'intérêt d'un usage de l'outil dans la version que nous avons utilisée. Il est en effet difficile, pour un non informaticien, de prendre en main la plate-forme de TAL pour procéder à des modifications de la chaîne mise en place. C'est pourquoi, une solution plus souple a été proposée. Une nouvelle version permet désormais d'exploiter LinguaStream sous la forme d'un service Web[11]. Celui-ci permet d'exploiter des chaînes d'analyse déjà créées, en ayant une interface simplifiée pour le réglage de quelques paramètres seulement – ceux choisis par l'informaticien qui met la chaîne d'analyse en ligne. En particulier, le texte en entrée, les lexiques, etc. peuvent être fournis par l'utilisateur pour tester une même chaîne sur différents corpus, avec différentes ressources.

[11] Cette version Web est accessible à l'adresse http://ls-web.info.unicaen.fr/, mais reste encore en développement à l'heure où nous bouclons cet article.

L'interaction entre linguistes et informaticiens reste bien sûr nécessaire, ne serait-ce que pour la mise en place des chaînes d'analyse et des choix des éléments qui feront l'objet d'un paramétrage.

Cet article avait aussi pour objectif de montrer la faisabilité d'une implémentation d'une grammaire locale, valable pour un jeu de langage particulier. Notre intention pour la suite est la projection de cette grammaire sur d'autres corpus, non pas à des fins d'analyse automatique, mais, plus modestement, d'exploration de corpus et de détermination d'indices pour l'interprétation. Plus précisément, et pour donner un exemple concret, nous travaillons actuellement sur l'analyse des lettres de lecteurs, envoyées à E. Sue pendant la publication des *Mystères de Paris* dans *le Journal des Débats*[12]. Les lecteurs de l'époque exprimaient dans ces lettres leurs avis sur le déroulement de l'histoire, leurs attentes, le style de l'auteur, leurs impressions générales, etc. Bref, ils faisaient part de leur expérience de lecteur, expérience dont l'analyse est utile pour les études sur la réception des œuvres littéraires. Le travail que nous menons est seulement amorcé, mais LinguaStream nous donne déjà la possibilité d'explorer ce corpus conséquent et de faciliter une analyse comparative entre les valeurs lectoriales et les formes linguistiques mobilisées par les lecteurs du 21e siècle, et celles exprimées, dans ce corpus, par les lecteurs du 19e. Nous espérons que ce travail, une fois mené à terme, montrera à son tour la pertinence d'une linguistique outillée.

Bouquet, S. (1999), « De la méthode directe aux investigations philosophiques de Wittgenstein. Savoirs et transferts de savoirs », in Langage et Société, ° 87 :41-77.

Charaudeau, P. (1993), « La critique cinématographique : faire voir et faire parler », La presse : produit, production, réception, Coll. Langages Discours et Sociétés, Paris, Didier Érudition : 47-70.

Covington, M. A. (1994), «GULP 3.1 : An Extension of Prolog for Unification-Based Grammar », Research Report AI-1994-06, Artificial Intelligence Center, The University of Georgia, Athens, Georgia, U.S.A.

Dufays, J.L., (2000), « Lire, c'est aussi évaluer. Autopsie des modes de jugement à l'œuvre dans diverses situations de lecture », in Études de linguistique appliquée, 119, 277-290.

Enjalbert, P. (éd.) (2005), *Sémantique et traitement automatique du langage naturel*, Paris, Lavoisier, Hermès Sciences.

Ferrari, S., Bilhaut, F., Widlöcher, A., Laignelet, M. (2005), « Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à

[12] Le corpus a été publié par Galvan, (1998).

l'évaluation de la détection automatique de cadres temporels ». In Williams, G. (éd.) Actes des 4èmes Journées de la Linguistique de Corpus, Lorient, France.

Galvan J.P., (1998), Les Mystères de Paris : Eugène Sue et ses lecteurs, Paris, L'Harmattan, (2 volumes).

Gross, M. (1995), « Une grammaire locale de l'expression des sentiments », Langue Française, 105 : 70-87.

Habert, B., Fabre, C., Issac, F. (1998), De l'écrit au numérique : constituer, documenter, normaliser un corpus électronique, Paris, InterEditions.

Hopper, P. (1998), « Emergent Grammar » in Tomasello, M., The new psychology of language cognitive and functional approaches to language structure, Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey London : 155-177.

Hunston, S. et Sinclair, J. (2000), « A Local Grammar of Evaluation », in Hunston, S. et Thompson, G. (eds), Evaluation in Text. Authorial Stance and the Construction of Discourse, Oxford University Press.

Legallois : 2007 « Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons: la grammaire de construction et la grammaire contextualiste », in Cahiers de l'Institut de Linguistique de Louvain (CILL), 31.2-4 : 109-127.

Legallois, D. & François, J. (2006) « Autour des grammaires de construction et de patterns » in Cahier du CRISCO n° 21 (publication accessible en ligne http://elsap1.unicaen.fr/).

Legallois, D., & Poudat, C. (2008), « Comment parler des livres que l'on a lus ? Discours et axiologie des avis des internautes », Semen 26 : 49-80

Martin, J., White, P. (2005), The Language of Evaluation: Appraisal in English, Basingstoke, UK, Palgrave Macmillan Hardcover.

Poibeau, Th. (2003), Extraction automatique d'information, Paris, Hermès

Schmid, H. (1994), « Probabilistic Part-of-Speech Tagging Using Decision Trees », Proceedings of the International Conference on New Methods in Language Processing. Manchester, U.K.

Vernier M., Mathet Y., Rioult F., Charnois T., Ferrari S. et Legallois D., (2007), « Classification de textes d'opinions : une approche mixte n-grammes et sémantique », in Actes de DEFT07 3ème DÉfi Fouille de Textes. 3 juillet 2007, Grenoble, France.

Vernier, M., Ferrari, S. et Legallois, D., (2007), « Discours évaluatif et suivi d'opinion », in Actes des 5èmes Journées de la Linguistique de Corpus, Lorient, France.

Widlöcher A., Bilhaut F., (2005), « La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus », in Jardino, M. (éd.), Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN 2005), Dourdan, France, ATALA, 2005.

Wittgenstein, L. (1961), Tractatus logico-philosophicus, *suivi de* Investigations philosophiques, Paris, Gallimard.

Wittgenstein, L. (1976), De la certitude, Paris, Gallimard.