Lecture Textométrique Différentielle (LTD) de textes législatifs comparables de l'Union européenne

Stéphane Patin[1], Maria Zimina[1], Serge Fleury[2]

[1]CLILLAC-ARP EA 3967, Université Paris Diderot-Paris 7, Paris, France
[2]CLESTHIA-EA 7345, Sorbonne Nouvelle-Paris 3, Paris, France

Abstract

Differential Textometric Reading (DTR) displays the results of *characteristic elements computation* (Lebart and Salem, 1994) for comparative reading of texts. Characteristic elements computed on multiple text annotation levels are highlighted on the screen to achieve visual differentiation of compared texts. We used DTR to contrast discourse features of legal French text corpora collected within the international project *Eurolect Observatory. Interlingual and intralingual analysis of EU legal varieties*. The suggested approach allowed fine-grained contrastive analysis of EU legal discourse (pragmatically vague and inclusive) and national legal discourse (more precise, circumscribed by French Legal System).

Résumé

La *Lecture Textométrique Différentielle* (LTD) montre au fil de la lecture comparative de textes leurs différences, ou leurs similarités caractéristiques, calculées par la méthode des *spécificités* (Lebart et Salem, 1994). Les éléments caractéristiques calculés sur annotations multiples sont rendus « visibles » au fil des textes par un système de surlignage. Appliquée aux textes législatifs comparables en français, collectés dans le cadre du projet international *Observatoire sur l'eurolecte. Analyse interlinguistique et intralinguistique des variétés juridiques dans l'Union européenne*, la LTD a permis de contraster deux schémas énonciatifs : celui de l'instance juridique communautaire, à vocation holistique mais doté d'un vocabulaire vague, produisant un cadre juridique adaptable à la législation nationale, et celui du système juridique français, beaucoup plus circonscrit et précis.

Key words: annotations, eurolecte, Lecture Textométrique Différentielle, méthode des spécificités, segments répétés, variétés juridiques

1. Lecture Textométrique Différentielle (LTD)

Nous appelons *Lecture Textométrique Différentielle* (LTD) une stratégie de lecture de textes comparables appuyée par l'affichage synchrone des résultats de leur analyse textométrique parallèle (différences ou similarités caractéristiques). Les deux ensembles textuels sont affichés simultanément à l'écran pour faciliter les comparaisons. La différentiation mobilise le calcul des *spécificités* (Lebart et Salem, 1994) : les éléments caractéristiques sélectionnés par seuillage dans chacun des ensembles textuels comparés sont rendus « visibles » au fil des textes par un système de surlignage (cf. *Figure 1*). La LTD est enrichie par la projection des résultats issus du calcul des répétitions segmentales sur annotations multiples (Zimina et Fleury, 2015) afin de garantir une prise en main « globale » de la différentiation à plusieurs niveaux d'analyse linguistique.

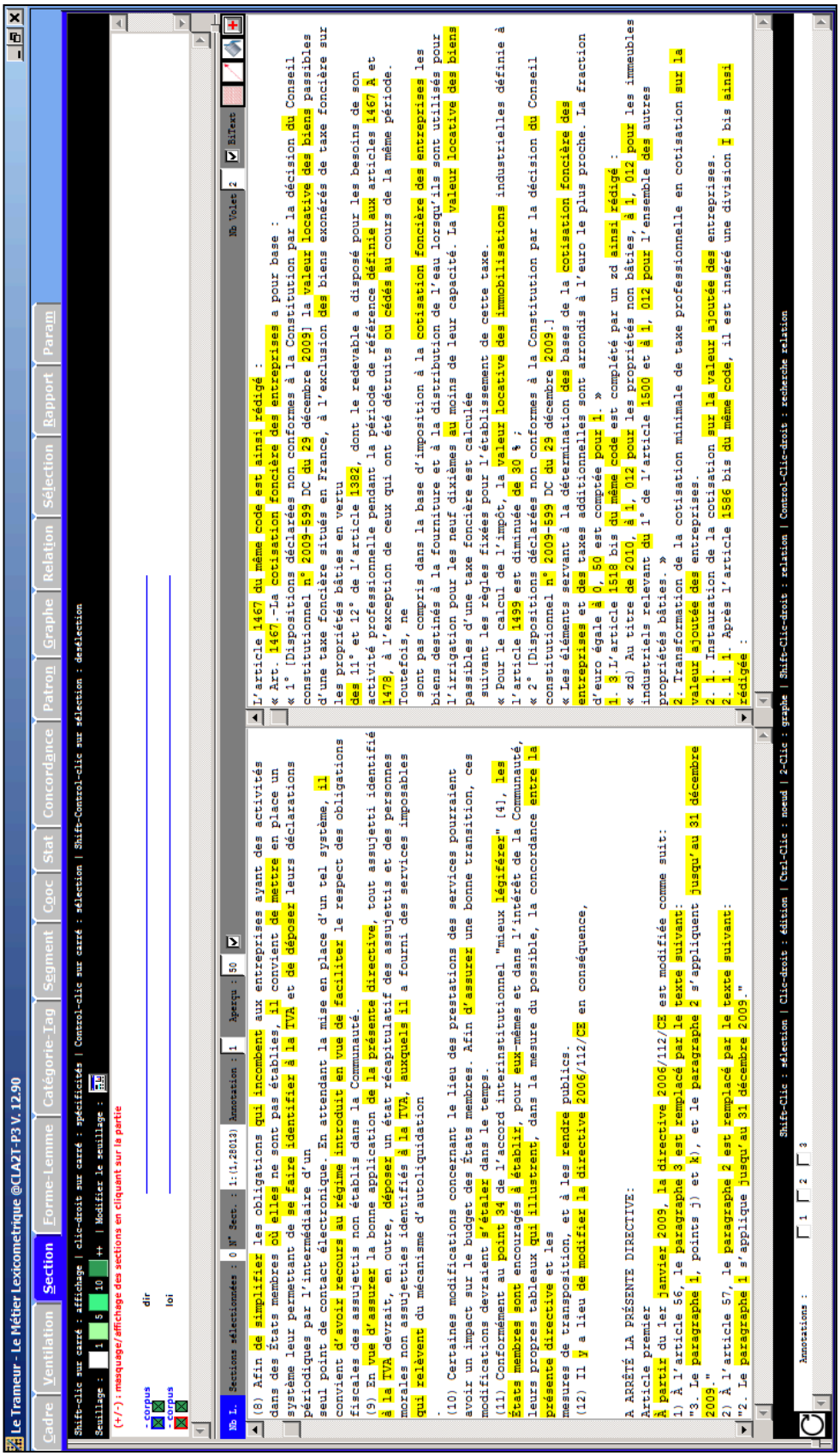

Figure 1 : Lecture Textométrique Différentielle (LTD) de textes juridiques comparables (directives européennes / lois nationales) avec surlignage des occurrences d'items spécifiques de chaque volet

Le recours à la LTD comme outil méthodologique peut avoir plusieurs objectifs. Elle peut aider à cerner un phénomène par l'examen des différentes causes, facteurs et processus qui y sont liés. Dans notre expérimentation, la LTD appliquée à des textes législatifs issus du droit communautaire et du droit national permet une meilleure compréhension des variétés juridiques.

2. Langue communautaire européenne

L'analyse de la caractérisation de la langue communautaire européenne suscite un intérêt grandissant parmi les linguistes, traducteurs et juristes comme en témoignent non seulement la littérature de plus en plus abondante à ce sujet (Gontrand, 1991 ; Goffin, 1994, 2002, 2005 ; Ciostek, 2014 ; Biel, 2014, etc.) mais aussi les différentes désignations néologiques employées pour décrire le phénomène. D'ailleurs si les termes tels que *eurolangue* ou *eurolecte* s'avèrent relativement neutres, d'autres, tels que l'anglais *eurofog*, le français *eurobabillage* ou encore l'allemand *euronebel*, ne manquent pas de signaler le caractère fortement cryptique de cette « langue » souvent perçue comme embrouillée.

La présente étude participe à ce débat en s'inscrivant dans le cadre du projet international *Observatoire sur l'eurolecte. Analyse interlinguistique et intralinguistique des variétés juridiques dans l'Union européenne* de la Faculté de traduction et d'interprétation de Rome.[1] Le projet cherche à déterminer les principales propriétés eurolectales à partir d'un corpus multilingue comparable composé de 660 directives de l'Union européenne de 1998 à 2008 (Corpus A de 29 398 122 occurrences), et de leurs transpositions dans 11 pays de l'UE, (Corpus B de 33 859 056 occurrences) : Allemagne, Angleterre, Espagne, Finlande, France, Grèce, Hollande, Italie, Lituanie, Malte et Pologne (cf. le *Tableau 1*) :

Langue	Directives (Corpus A)		Lois nationales (Corpus B)		TOTAL occ.
	Nb directives	Nb occurrences	Nb directives	Nb occurrences	
Anglais	660	3 700 533	674	8 732 916	12 433 449
Allemand	660	3 348 510	467	3 067 711	64 16 221
Espagnol	660	4 140 609	469	5 929 495	10 070 104
Français	660	4 181 496	129	1 862 241	6 043 737
Italien	660	3 749 550	277	2 837 488	6 587 038
Néerlandais	660	3 701 842	504	4 012 239	7 714 081
Polonais	658	3 422 437	482	6 528 417	9 950 854
Maltais	656	3 364 329	276	834 086	4 198 415
TOTAL occ.		**29 398 122**		**33 859 056**	**63 257 178**

Tableau 1 : Corpus A et B. Observatoire sur l'eurolecte. Analyse interlinguistique et intralinguistique des variétés juridiques dans l'Union européenne[2]

[1] Le projet *Observatoire sur l'eurolecte. Analyse interlinguistique et intralinguistiques des variétés juridiques dans l'Union européenne* dirigé par Laura Mori regroupe 11 juri-linguistes et linguistes dont Stéphane Patin. Une présentation détaillée du projet est disponible en ligne : http://www.unint.eu/en/research/research-groups/39-higher-education/490-eurolect-observatory-interlingual-and-intralingual-analysis-of-legal-varieties-in-the-eu-setting.html

[2] Les résultats statistiques pour les données en français s'appuient sur la segmentation en occurrences réalisée à l'aide du logiciel *Le Trameur* (Fleury et Zimina, 2014).

La constitution de ces corpus, au format XML, structurés par des balises délimitant quatre parties (titre, préambule, dispositif et annexe), répond à plusieurs objectifs. Sur le plan linguistique et traductologique, il s'agit de décrire les variétés juridiques qui se sont développées dans l'Union européenne, à travers l'examen des diasystèmes linguistiques d'un certain nombre d'États membres, et mettre, ainsi, en évidence les différences linguistiques entre le système linguistique employé dans les directives et les variétés juridiques nationales. Sur le plan pragmatique et institutionnel, le projet cherche à mesurer la lisibilité des textes appartenant aux différents corpus, à fournir des données aux Services linguistiques des institutions de l'Union européenne, aux Parlements nationaux et aux Assemblées législatives régionales ou autonomes, et à produire des résultats utiles à la rédaction des textes législatifs européens et nationaux.

3. Corpus de travail

Notre travail prend appui sur le volet français du corpus comparable multilingue de l'*Observatoire sur l'eurolecte* présenté dans le *Tableau 1*. Ce volet est constitué des 660 directives de l'Union européenne de 1998 à 2008 (Corpus A) et de leur transposition en France, dans 129 textes de lois et ordonnances (Corpus B), comptabilisant, respectivement, 4 181 496 (***FR-A***) et 1 862 241 occurrences (***FR-B***). La présente étude cherche à localiser, quantifier et interpréter les différences « textuelles » entre les deux corpus afin de mettre en lumière les propriétés de l'eurolecte. Pour y parvenir, on déploie la stratégie de la LTD avec *Le Trameur* (Fleury et Zimina, 2014).

4. Principes de différentiation

4.1. Sélection des occurrences d'items spécifiques d'une partie du corpus

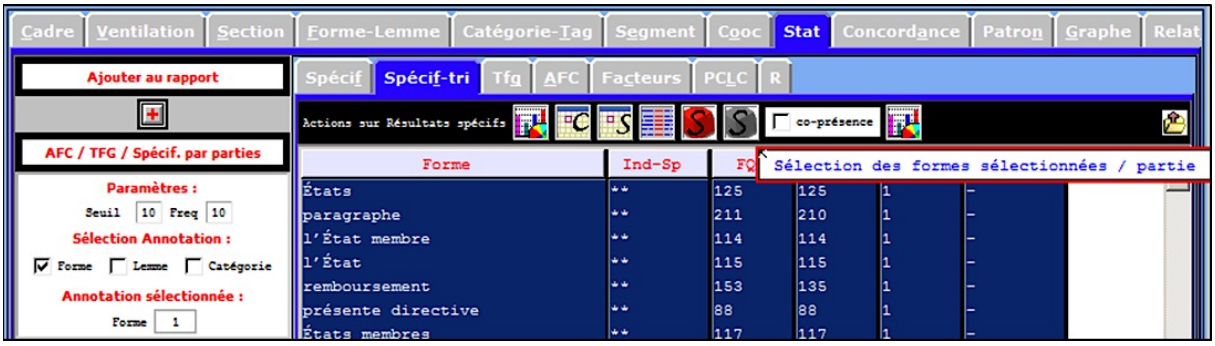

Figure 2 : Sélection différentielle par partie des occurrences d'items spécifiques

FR-A et ***FR-B*** sont analysés en termes de *spécificités* (Lebart et Salem, 1994) avec *Le Trameur*. Les diagnostics de spécificités sont présentés sous forme de listes d'*items* (par exemple, formes et segments répétés) que l'on peut trier par la valeur d'*indice de spécificité*. A partir de cette liste, il est possible de réaliser une sélection automatique d'*items* caractéristiques de chaque partie du corpus pour mener des explorations ultérieures (analyse des contextes, concordances, graphiques, cartographies du corpus, etc.). Afin de cibler uniquement les occurrences de la partie où le diagnostic de spécificité a été posé en LTD, les *Sélections*[3] d'*items* spécifiques sont toujours réalisées avec filtrage. Cette restriction est un

[3] Les *Sélections* sont des items correspondant aux occurrences d'un type : forme, lemme, patron morphosyntaxique, expression régulière croisant plusieurs annotations (Zimina et Fleury, 2014 ; 2015).

prérequis pour la mise en route de la différentiation réalisée grâce aux principes de l'architecture *Trame/Cadre* implémentée dans *Le Trameur* (Zimina et Fleury, 2015).

La *Figure 2* montre l'accès à la *Sélection* différentielle à partir d'une liste de spécificités par partie. Les positions des occurrences sélectionnées dans l'extrait de ***FR-A*** (directives de l'Union européenne) sont mémorisées par le gestionnaire de *Sélection* (cf. *Figure 3*). Les occurrences concernées par la *Sélection* sont automatiquement surlignées dans le texte. Par exemple, sur la *Figure 4*, la *Sélection* surlignée en jaune matérialise les formes et segments répétés dont la valeur d'*indice de spécificité* est supérieure ou égale à 20. Le seuillage peut varier au fil de la lecture en fonction du cadre d'exploration.

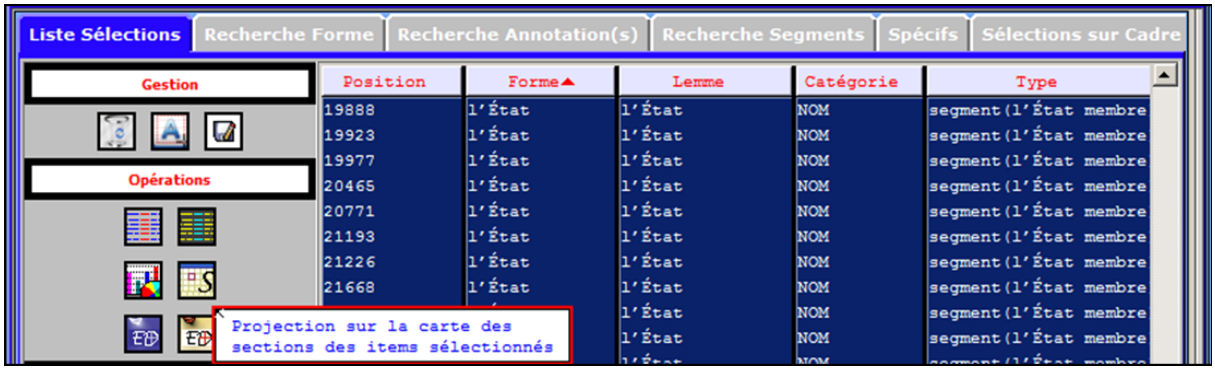

Figure 3 : Localisation des occurrences sélectionnées en corpus via le gestionnaire de Sélection

4.2. Spécificités et Sélections sur annotations multiples

La LTD tient compte de plusieurs niveaux d'annotation du corpus. Après l'étiquetage du corpus de travail par *TreeTagger*[4], les *spécificités* morpho-syntaxiques (catégories et figements de catégories caractéristiques) sont utilisées pour donner un éclairage supplémentaire aux données analysées. Cette approche outillée nous permet d'apprécier les emplois spécifiques à chaque niveau d'annotation, par exemple l'emploi caractéristique du conditionnel dans les directives (cf. *Figure 5*).

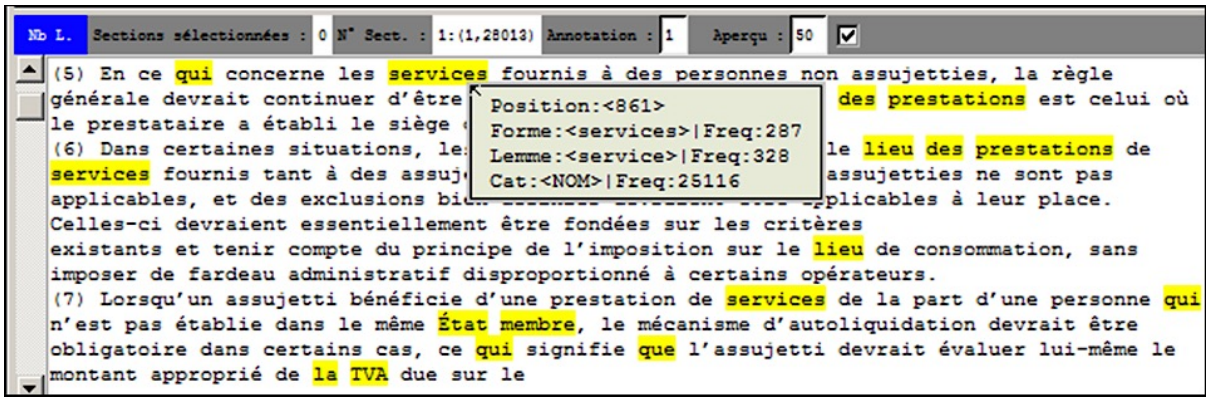

*Figure 4 : Lecture Textométrique Différentielle (LTD) sur l'extrait **FR-A** (directives européennes) avec marquage des spécificités lexicales (formes et segments répétés)*

[4] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

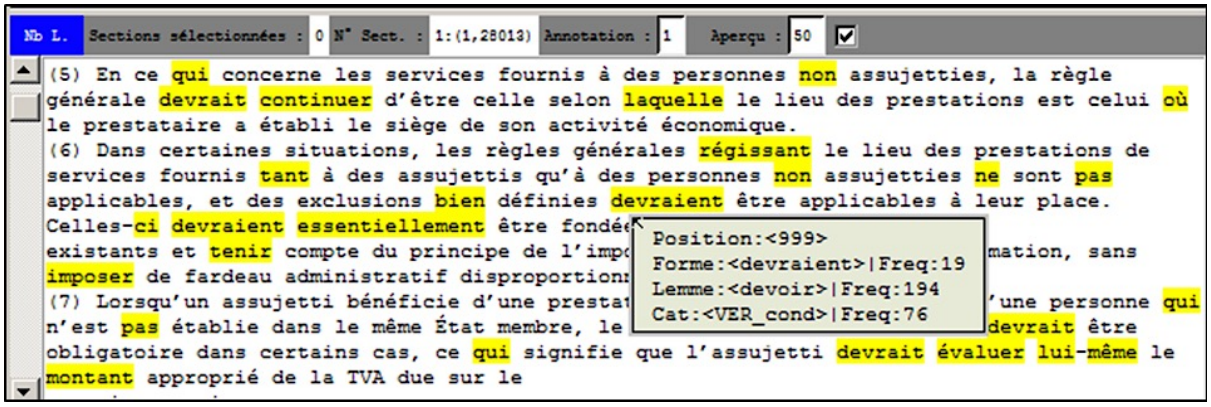

*Figure 5 : Lecture Textométrique Différentielle (LTD) sur l'extrait **FR-A** (directives européennes) avec marquage des spécificités morpho-syntaxiques*

4.3. Visualisation synchrone avec points d'ancrage comme aide à l'interprétation

On peut orienter la LTD vers des passages textuels relevant d'une concentration de traits linguistiques saillants qui mobilisent plusieurs niveaux d'annotation (formes, lemmes, catégories morphosyntaxiques, etc.). Cette *Sélection* de points d'ancrage oriente l'attention vers des contextes particuliers en facilitant l'analyse différentielle.

Sur la *Figure 6*, ce sont les occurrences du terme « prestations » surlignées en rouge qui servent de points d'ancrage. Les deux ensembles textuels sont présentés sous forme de bi-texte. Le premier volet à gauche correspond aux directives européennes (***FR-A***) tandis que le deuxième à droite montre la transposition de ces directives dans la législation française (***FR-B***). Les deux volets sont mis en correspondance dans *Le Trameur* à l'aide des fonctions de visualisation de l'appariement des *sections* (contextes délimités) en correspondance. L'alignement est réalisé sur la base de la correspondance *directive => transposition*.

L'alignement présenté sur la *Figure 6* affiche un extrait de la *Directive 2008/8/CE du Conseil du 12 février 2008 modifiant la directive 2006/112/CE en ce qui concerne le lieu des prestations de services*, et une transposition dans le texte de la *LOI n° 2009-1673 du 30 décembre 2009 de finances pour 2010 (1) NOR: BCFX0921637L.*

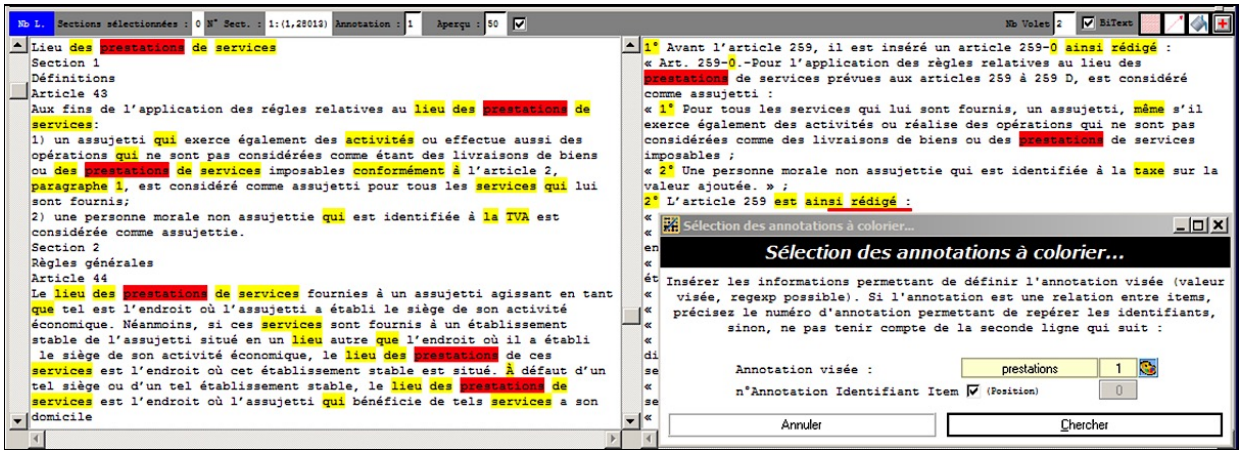

Figure 6 : Lecture Textométrique Différentielle (LTD) avec visualisation synchrone des extraits comparables relevant de la concentration des occurrences du terme « prestations »

5. Interprétation des résultats

Au cours de l'étude, la LTD a permis une contextualisation simultanée des spécificités du corpus à plusieurs niveaux d'analyse linguistique. Si les concordances offrent des accès contextuels ciblés autours de certains phénomènes textuels, ils ne permettent pas en revanche une prise en compte globale de la différentiation au fil de la lecture. De même pour les diagrammes de spécificités et des représentations de diagnostiques sous forme de listes. De ce point de vue, la LTD a constitué un complément utile aux schémas d'exploration habituels du corpus de travail.

A la lumière de ces analyses contrastives, des tendances discursives révélées par l'étude des *spécificités* via la LTD, amènent à penser que l'instance juridique communautaire prend en charge une énonciation à vocation holistique mais qui emploie un vocabulaire assez généraliste, octroyant ainsi aux directives qu'elle produit un cadre juridique suffisamment large pour qu'elles puissent être adaptées au système juridique national, alors que ce dernier réalise une énonciation plus précise et circonscrite à l'ordonnancement français.

5.1. Énonciation totalisante et imprécise de l'instance juridique communautaire

Les directives relèvent d'une énonciation juridique communautaire imprécise mais à vocation exhaustive. Au niveau syntaxique, on observe la construction de phrases complexes avec abondance de subordonnants tels que les pronoms relatifs, les conjonctions de subordination exprimant l'éventualité, l'alternative, la condition du type « si » ou l'emploi caractéristique de l'adjectif indéfini « tout » (cf. *Figures 7-8*). C'est à cette fin que l'on peut également interpréter le recours spécifique au conditionnel (cf. *Figure 9*).

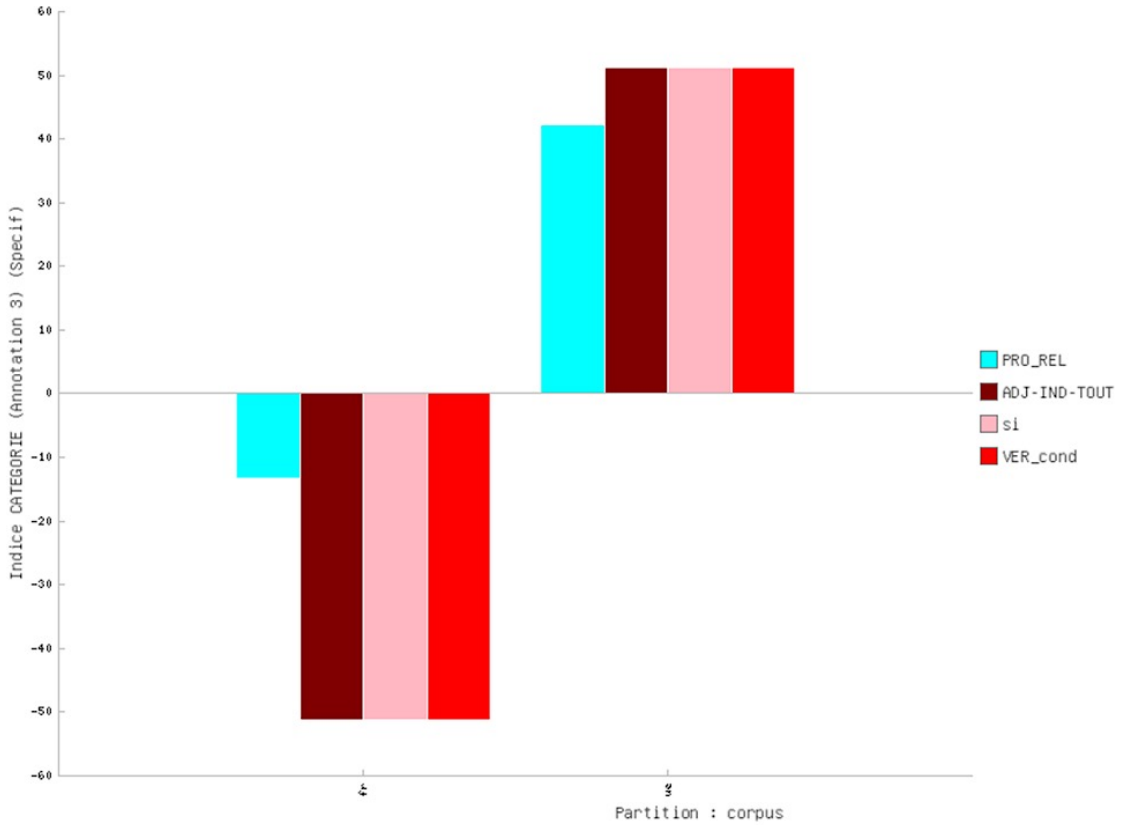

Figure 7 : Diagramme de spécificités multiannotation généré par Le Trameur (pronoms relatifs, verbes au conditionnel, l'adjectif indéfini « tout » et la conjonction de subordination « si »)

toute	personne qui importe dans la Communauté des
toutes	les directives pertinentes afin d'éviter que
tout	équipement qui est soit un "équipement
toute	interférence qui compromet le fonctionnement d'un
tous	les appareils: a) la protection de
toute	autre personne, y compris les objectifs,
tous	les services fournis par l'interface correspondante
toutes	les informations nécessaires pour permettre aux fabricants

Figure 8 : Extrait de la concordance de l'adjectif indéfini « tout » (expansion à droite dans les directives européennes)

ire les charges administratives pour les exploitants d'aéronefs, il	serait	souhaitable que chaque exploitant d'aéronef relève d'un État m
xploitant d'aéronef relève d'un État membre. Les États membres	devraient	être tenus de veiller à ce que les exploitants
l'État membre responsable de l'y contraindre, les États membres	devraient	agir solidairement. En conséquence, il convient que, en
itement équitable des exploitants d'aéronefs, les États membres	devraient	suivre des règles harmonisées pour l'administration des exploita

Figure 9 : Extrait de la concordance des verbes au conditionnel dans les directives européennes

Sur le plan lexical, les directives européennes contiennent spécifiquement des lexèmes fonctionnant comme des hyperonymes tels que la désignation des destinataires : « État membre », « autorités compétentes » (cf. *Figure 10*).

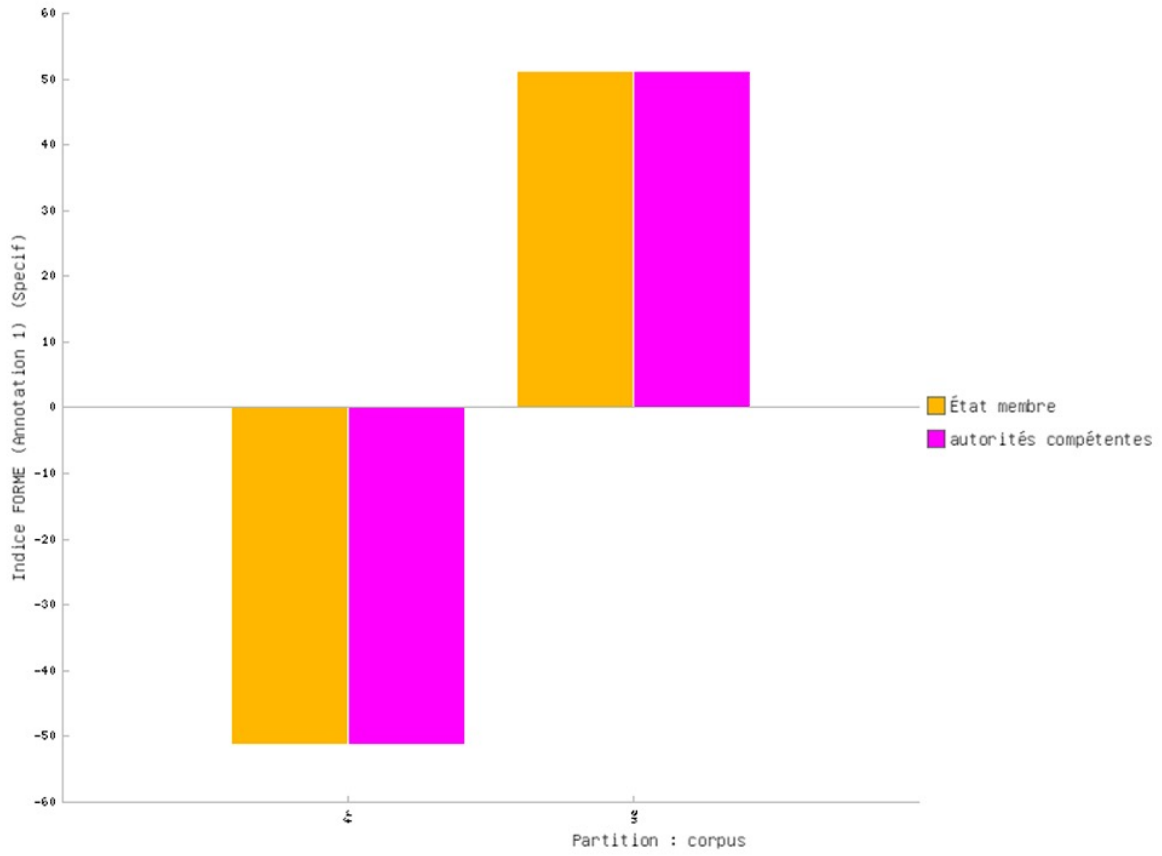

Figure 10 : Diagramme de spécificités des lexèmes « État membre » et « autorités compétentes »

C'est en contexte, au fil de la LTD, que l'on parvient à cerner les interactions discursives entre ces éléments caractéristiques des directives. Sur les captures d'écran présentées dans la *Figure 11*, les spécificités relevées au niveau des catégories morpho-syntaxiques sont surlignées en rouge tandis que les formes et segments répétés ayant déclenché le diagnostic de spécificité sont surlignés en jaune. Cette visualisation des *spécificités multiannotation* s'appuie sur le modèle de données *Trame/Cadre* qui rend possible la prise en compte simultanée de plusieurs couches d'annotation de corpus lors de l'analyse textométrique (Zimina et Fleury, 2014 ; 2015).

Figure 11 : LTD des directives européennes par spécificités multiannotation (catégories morpho-syntaxiques en rouge, formes et segments en jaune)

5.2. Énonciation précise du cadre juridique national

La transposition des directives implique une énonciation plus précise, reflet d'une adaptation juridique, comme le montre l'extrait sur la *Figure 12*. Le processus de transposition figure formellement dans le texte national. Il passe par la modification, l'insertion et la suppression d'éléments comme en témoigne la surreprésentation des segments « ainsi rédigés », « il est inséré », « est inséré un article », « complété par un alinéa », « sont supprimés », « est supprimé » (indice de spécificité supérieur ou égal à 20).

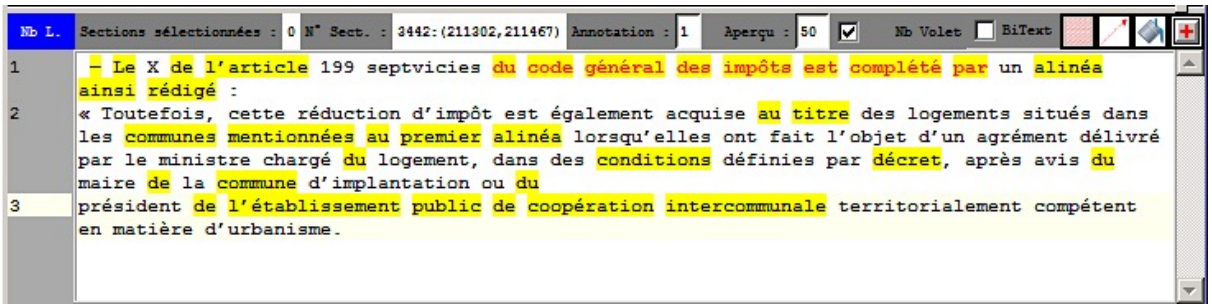

*Figure 12 : LTD d'un extrait de la législation nationale (**FR-B**) à l'aide des spécificités lexicales*

Cette transposition se fait dans un cadre juridique précis où la « directive » devient « loi » moyennant un « décret » (« par décret », « en décret », etc.), où « autorités compétentes », « État membre » deviennent, par exemple, « le Gouvernement » (cf. *Figure 13*).

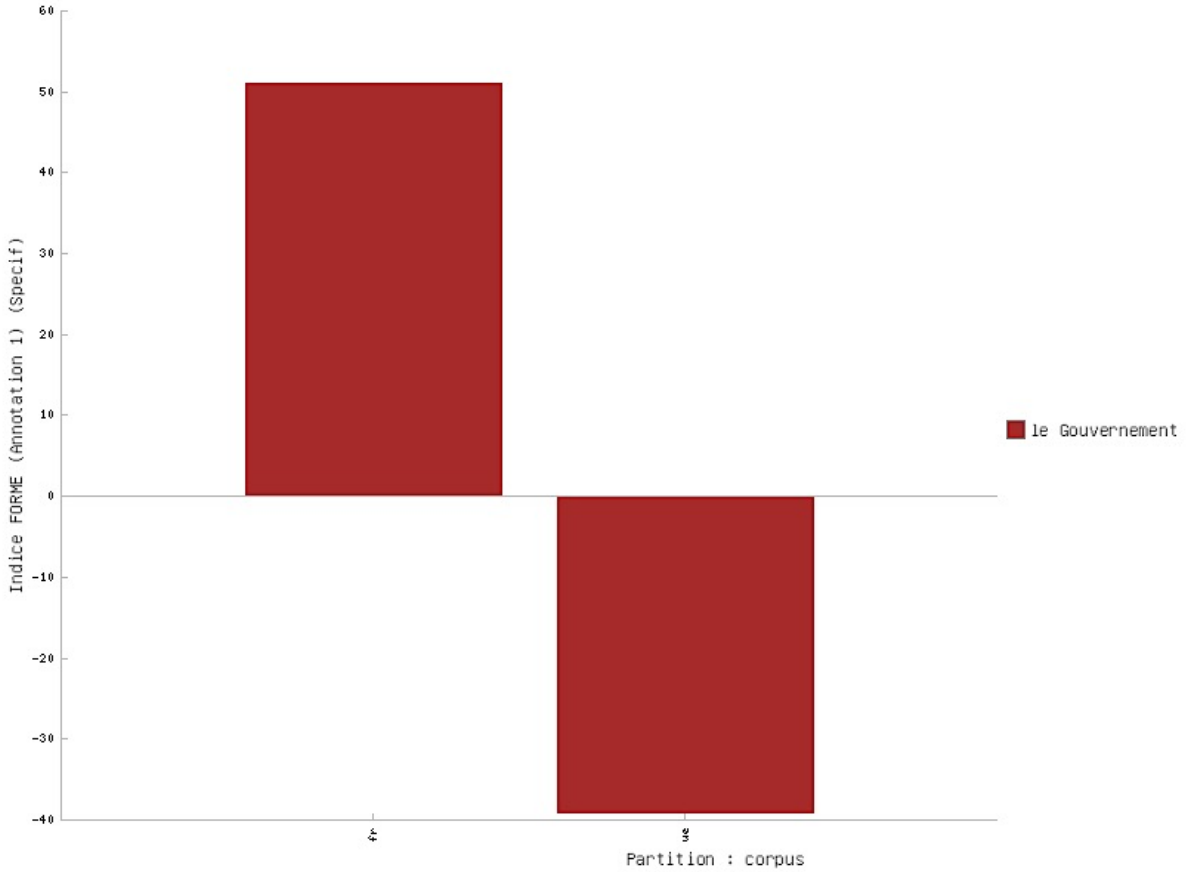

Figure 13 : Diagramme de spécificités du segment répété « le Gouvernement »

6. Conclusion et perspectives

Nous avons montré comment la *Lecture Textométrique Différentielle* (LTD) peut contribuer à l'analyse de textes juridiques comparables en montrant au fil de la lecture leurs différences ou leurs similarités caractéristiques calculées par la méthode des *spécificités* (Lebart et Salem, 1994). Cette approche crée une passerelle entre les résultats des calculs statistiques et l'interprétation du contexte au fil de la lecture. Sur le plan discursif, il devient possible d'observer les liens entre plusieurs phénomènes attestés sur le plan quantitatif (répétitions segmentales, figements des catégories morpho-syntaxiques, surreprésentation (ou sous-représentation) d'*items* avec prises en compte d'annotations multiples, etc. La LTD se prête bien à l'analyse contrastive de textes qui autorisent des comparaisons. Elle peut être utilisée pour l'affichage synchrone des passages en correspondance (traductions, réécritures, paraphrases, versions, etc.). Dans notre expérience, cette stratégie a été mobilisée pour contraster deux schémas énonciatifs : celui de l'instance juridique communautaire, totalisante mais floue, produisant un cadre juridique adaptable à la législation nationale, et celui du système juridique français, beaucoup plus précis et circonscrit.

La LTD fait partie des fonctionnalités du logiciel de textométrie *Le Trameur*.[5] Son activation se réalise via le gestionnaire de *Sélections*, elle se matérialise via l'éditeur de la *carte des*

[5] http://www.tal.univ-paris3.fr/trameur/

sections (représentation cartographique de la *Trame* du corpus) avec affichage synchrone des *sections* en correspondance. Nous envisageons plusieurs pistes de travail afin d'offrir plus d'options à l'utilisateur au fil de la LTD. A terme, il serait possible d'intégrer un *seuillage visuel* : la distinction des seuils de spécificités dans les options de surlignage au fil du texte (du plus foncé au plus clair, compte tenu de niveaux d'*indice de spécificité*).

Sur le plan d'exploration de corpus de textes législatifs comparables ***FR-A/FR-B***, une des pistes consiste à réfléchir à la représentation visuelle de *sections* en correspondance en cas de liens type « un-pour-plusieurs » (une directive européenne peut avoir plusieurs transpositions dans différentes lois nationales). Dans ce cas, on peut envisager des alignements multiples, à l'instar de l'analyse contrastive de plusieurs traductions d'un même texte (Zimina et Fleury, 2015).

Compte tenu des premiers résultats expérimentaux et des chantiers en perspective, nous sommes convaincus que la LTD peut s'intégrer avec succès à la pratique de l'analyse textométrique sur corpus. Elle peut constituer un complément utile aux diagrammes, dictionnaires, concordances et autres aides interprétatives qui permettent à l'utilisateur d'interagir avec les corpus informatisés.

Références

Biel L. (2014). *Lost in the Eurofog. The Textual Fit of Translated Law*. Frankfurt am Mein: Peter Lang.

Ciostek A. (2014). "Eurolangue et sa productivité : quelques tendances." In *Roczniki Humanistyczne*, vol. 62 (8): Corpus Linguistics and Translation Studies, Lublin: The Learned Society of KUL & John Paul II Catholic University of Lublin, pages 65-77.

Fischer M. (2010). "Language (policy), translation and terminology in the European Union." In Marcel Thelen, M. and Steurs F. editors, *Terminology in Everyday Life*. Amsterdam: Benjamins, pages 21-33.

Fleury S. and Zimina M. (2014). "Trameur: A Framework for Annotated Text Corpora Exploration." *Proc. of COLING 2014 (25th International Conference on Computational Linguistics)*: System Demonstrations, August 2014, Dublin, Ireland, pages 57-61.

Gontrand F. (1991). *Parlez-vous eurocrate ? Les 1000 mots clés du Marché Unique.* Les Éditions d'Organisation, Paris.

Goffin R. (1994). "L'Eurolecte : oui, jargon communautaire : non." *Meta*, Hommage à B. Quemada, vol.39(4), pages 636-642.

Goffin R. (2002). "Eurolecte : Analyse contrastive de quinze eurolexies néologiques." *Cahiers de lexicologie*, vol.1(80), pages 167-177.

Goffin, R. (2005). "Quels corpus et quelles approches pour une description contrastive de l'eurolecte?" Mots, termes et contextes. *7èmes Journées scientifiques du réseau de chercheurs Lexicologie, Terminologie, Traduction*, 8-10 septembre 2005, Bruxelles, Belgique.

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

Zimina M. et Fleury S. (2014). "Approche systémique de la résonance textuelle multilingue." *Actes des 12es Journées internationales d'Analyse statistique des Données Textuelles*, juin 2014, Paris, pages 717-728.

Zimina M. et Fleury S. (2015). "Perspectives de l'architecture *Trame/Cadre* pour les alignements multilingues. 2010. *Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles*, vol.11(1).