

# Approche systémique de la résonance textuelle multilingue

Maria Zimina<sup>1</sup>, Serge Fleury<sup>2</sup>

<sup>1</sup> Université Paris Diderot – maria.zimina@eila.univ-paris-diderot.fr

<sup>2</sup> Université Sorbonne Nouvelle – serge.fleury@univ-paris3.fr

## Abstract

In a bi-text, corresponding textometric *Threads* are revealed by both lexical correspondences and functional equivalence of related linguistic features. Following this text alignment representation, we develop a systemic approach of *multilingual textual resonance*, simultaneously covering several levels of linguistic analysis. This approach is presented through a series of textometric explorations of the comparable corpus BBC\_Lenta.RU. It is composed of BBC News (2001-2005) and their adaptations into Russian published by www.lenta.ru.

## Résumé

Dans un bi-texte, les *Trames* textométriques en correspondance sont matérialisées à la fois par des correspondances lexicales et des équivalences fonctionnelles de traits linguistiques corrélées. Pour en tenir compte au cours de l'alignement de textes, nous proposons une approche systémique de la *résonance textuelle multilingue* qui mobilise simultanément plusieurs niveaux d'analyse linguistique. L'approche est présentée à l'aide d'une série d'explorations textométriques du corpus comparable BBC\_Lenta.RU. Il est composé de textes d'actualités diffusés en anglais par la chaîne britannique BBC entre 2001 et 2005, et de leurs traductions-adaptations en russe publiées par le site web Lenta.ru.

**Mots-clés :** textométrie, traduction, linguistique systémique fonctionnelle, résonance textuelle multilingue

## 1. Introduction

L'analyse automatique de corpus multilingues parallèles et comparables a beaucoup progressé ces dernières années. Les avancées concernent aussi bien les couples de langues analysés, que les types de relations que l'on parvient à repérer entre les unités de bi-textes (Tiedemann, 2011). Le volume de ressources textuelles accumulé et la richesse de l'outillage méthodologique élaboré conduisent à une réflexion sur la systématisation des liens fonctionnels que l'on parvient à repérer dans le bi-texte.

Au-delà des listes de correspondances traductionnelles, le *bi-texte* est un ensemble cohérent d'éléments de plusieurs niveaux qui fonctionnent parallèlement. Ces éléments s'articulent selon des conventions lexicales, morphologiques, syntaxiques, etc. Sur le plan systémique, c'est la dynamique de la *lecture active* qui rend perceptible l'ensemble des processus coordonnés entre eux et synchronisés avec l'environnement pour la création de l'équivalence de sens (Plassard, 2007). Cette dynamique de lecture peut se construire autour de la navigation selon plusieurs parcours interactifs : mise en correspondance des éléments équivalents, recherche de schémas de traduction, etc. (*TransRead* – ANR, en cours).

## 2. Résonance textuelle multilingue

La lecture interactive du bi-texte peut être facilitée par les méthodes textométriques faisant appel à la *résonance textuelle*. Mobilisée d'abord en analyse de discours pour la mise en évidence de figements, reprises, paraphrases et réseaux textuels, la *résonance textuelle* s'est

progressivement étendue à l'étude de l'intertextualité interlangue en traduction (Salem, Fleury, éditeurs, 2009).

Le schéma de la *résonance textuelle* peut être résumé de la façon suivante (Salem, 2004) :

**Etape 1** : la mise en correspondance des ensembles textuels (textes, fragments, etc.) qui autorisent des comparaisons. 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.

**Etape 2** : la sélection d'un sous-ensemble d'unités dans un des volets du corpus pour amorcer le processus de résonance (*termes d'induction*).

**Etape 3** : la sélection induite sur les fragments du second volet du corpus sert de point de départ à la sélection des termes que leur présence (ou leur absence) rend caractéristiques pour cette dernière sélection (*termes induits*).

En cas de *correspondance partielle* entre les *termes d'induction* et les *termes induits*, le processus est réitéré dans la sélection où les *termes d'induction* sont présents et les *termes induits* absents (et vice versa) (Zimina, 2004 ; 2005).

Par exemple, dans un corpus parallèle de textes juridiques français/anglais alignés au niveau de la phrase, on peut amorcer la résonance textuelle en sélectionnant toutes les phrases du volet français qui contiennent les occurrences de la forme « requête » (c'est le *terme d'induction*). La sélection induite sur le fragment composé des phrases équivalentes dans le volet anglais fait ressortir la forme « petition » (c'est le *terme induit*). Les deux termes sont en correspondance partielle. Le terme « petition » est alors à nouveau utilisé pour amorcer la résonance. La sélection ne tient compte que des phrases alignées avec celles qui contiennent « petition » en anglais, sans que la forme « requête » y soit attestée. On induit ainsi un autre terme français « recours ».

Ce principe d'écrémage successif est mobilisé pour la recherche de liens multiples en corpus. Appuyée par la représentation topographique de la surface bi-textuelle, l'implémentation logicielle de ce principe a permis le développement des fonctions de navigation dynamique intertextuelle (Fleury, Zimina, 2008).

### 3. Approche systémique et textométrie

L'analyse textométrique mobilise naturellement plusieurs niveaux d'analyse linguistique et partage le modèle théorique élaboré dans le cadre de la *Linguistique Systémique Fonctionnelle* (désormais notée LSF) (Banks, 2005 ; Gledhill, 2011). La LSF fait intervenir conjointement différents niveaux d'analyse afin d'expliquer comment l'ensemble du matériel langagier est naturellement employé pour répondre aux besoins communicationnels de l'homme. Ainsi, la morphologie, la syntaxe et le stock lexical correspondent au même niveau d'analyse décrit par le terme *lexicogrammaire* (Halliday, 1961). Le caractère fonctionnel du *lexicogrammaire* est reflété par le lien entretenu par le texte avec son environnement (contexte).

La dimension fonctionnelle de la LSF permet de comprendre comment les différents composants de la langue fonctionnent ensemble et génèrent du sens dans un contexte social. En traduction vers une langue cible, cette analyse aide à restituer les fonctions d'un texte source en privilégiant des équivalences de traits linguistiques caractéristiques d'une situation de communication donnée (Manfredi, 2008). En textométrie, cette approche inductive se prête bien à la modélisation linguistique de la dynamique de la résonance textuelle multilingue :

à partir de « signaux » récoltés, on avance vers une construction systémique d'un modèle d'équivalences en corpus. La prise en compte de plusieurs niveaux fonctionnels qui constituent la bi-textualité aide à interpréter les phénomènes linguistiques observés statistiquement. Ils relèvent des équivalences matérialisées à la fois par la présence d'un vocabulaire évocateur et d'un ensemble de traits linguistiques qui réalisent le cadrage nécessaire au plan fonctionnel.

En effet, chaque texte d'un corpus bi-textuel existe dans une situation de communication donnée et véhicule un contenu dans un objectif qui peut être précisé en référence à son type et son genre. Par exemple, dans un corpus relevant du genre « actualités médiatiques », le « format » textuel utilisé pour la communication des actualités scientifiques sera différent de celui qui caractérise la présentation des informations sur les zones de conflits militaires. Les traits structurants de ces formats offrent des indices pour formaliser l'alignement textométrique des éléments textuels correspondants.

#### 4. Corpus *BBC\_LentaRU*

Une série d'explorations textométriques du corpus *BBC\_Lenta.RU* illustrent l'approche systémique de la résonance textuelle. Le corpus est composé de 2 345 textes correspondant à des fils d'actualités diffusés par la chaîne britannique *BBC* entre 2001 et 2005, et de leurs traductions-adaptations en russe publiées par *Lenta.ru*, un très populaire site d'actualités basé à Moscou. Les textes originaux en anglais comptent plus d'un million d'occurrences, tandis que leurs traductions-adaptations en russe sont, en moyenne, deux fois plus courtes (moins de 500 000 occurrences au total).

Récolté initialement pour une étude portant sur l'alignement d'entités nommées (Klementiev, Roth, 2006)<sup>1</sup>, ce corpus anglais/russe permet d'étudier des ensembles textuels comparables qui ne se prêtent pas à l'alignement formel au niveau des phrases ou paragraphes.

#### 5. Trames textométriques parallèles

##### 5.1. Construction de Trames

Le caractère informatif des textes médiatiques *BBC\_Lenta.RU* (écrits pour refléter l'actualité selon deux traditions journalistiques différentes) a plusieurs spécificités qui apparaissent à travers l'étude des schémas lexico-grammaticaux propres à ce type de texte. Les traits caractéristiques de ces schémas offrent des indices pour l'alignement des structures textuelles correspondantes. Le recensement automatique de ces indices est possible grâce à la prise en compte des annotations multiples sur les unités du texte au cours de l'analyse statistique.

Après la segmentation automatique de chaque volet du bi-texte, la numérotation des items découpés dans les textes de départ permet de constituer un système de coordonnées (un pour chaque volet) dans lequel chaque item est repéré par son numéro d'ordre : la *Trame* textométrique.

Plusieurs annotations textuelles peuvent se greffer sur une *Trame*. Ces annotations reflètent les états du corpus (lemmatisation, étiquetage automatique des catégories grammaticales, relations de dépendance syntaxique, analyse sémantique, etc.). Pour générer et gérer des annotations multiples sur les unités textuelles (Fleury, 2013), nous utilisons les fonctionnalités

---

<sup>1</sup> Le corpus est disponible sur le site <[http://cogcomp.cs.illinois.edu/page/resource\\_view/1](http://cogcomp.cs.illinois.edu/page/resource_view/1)> [février 2014]

du logiciel *Le Trameur*<sup>2</sup> qui intègre par défaut le programme *treetagger* : système d'étiquetage automatique des catégories grammaticales des mots avec lemmatisation (Schmid, 1994). Ce logiciel permet la gestion d'autres types d'annotations en parties du discours, comme nous le faisons dans ce travail.

Le volet anglais du corpus *BBC\_Lenta.RU* est annoté à l'aide du jeu d'étiquettes issu du projet *The Penn Treebank*.<sup>3</sup> L'étiquetage morphosyntaxique du volet russe utilise l'annotation développée pour le corpus national russe (*Национальный корпус русского языка*).<sup>4</sup>

Après étiquetage, le dénombrement des suites de catégories grammaticales qui se succèdent en corpus est réalisé par le calcul des répétitions segmentales, en mobilisant la méthode des *segments répétés* (Salem, 1987). Cette analyse statistique constitue une première étape de recensement automatique de traits caractéristiques de ce type de textes. Elle pourrait être suivie d'une phase de filtrage qui vise à éliminer des répétitions déjà incluses dans d'autres segments dénombrés.

Les figements recensés quantitativement résultent des emplois propres à chaque langue dans le contexte étudié. Ces dénombrements des catégories peuvent alors être analysés sur des bases typologiques. L'approche textométrique permet de préciser les types de rencontres entre les catégories et leurs fréquences en fonction de la distribution du stock lexical dans chaque volet du corpus. Les *Trames* textométriques se prêtent particulièrement bien à la modélisation linguistique de ces phénomènes en corpus.

## 5.2. Alignement de Trames par résonance textuelle

Pour aligner les *Trames* textométriques, on mobilise la résonance textuelle. L'induction est amorcée à l'aide d'une sélection par *seuillage* à partir d'une évaluation de la quantité des occurrences d'un terme (ou d'une liste de termes) dans les fragments textuels en correspondance. On amorce l'induction au niveau des occurrences de formes graphiques et on la propage vers le niveau d'annotation sélectionné (cf. figure 1).

Par exemple, dans le corpus *BBC\_Lenta.RU*, l'analyse des spécificités lexicales des textes russes qui correspondent à ceux qui contiennent « *people* » (2 601 occ.) dans le volet anglais, fait ressortir immédiatement sa principale équivalence contextuelle dans le volet russe : « *человек* » (813 occ.). Si l'équivalence « *people* » | « *человек* » (en français : *personnes*) n'a rien de surprenant, l'information flexionnelle concernant le terme russe est utile pour orienter la suite de l'analyse. À partir de la correspondance des termes « *people* » | « *человек* », la résonance est propagée à d'autres niveaux d'analyse linguistique. La sélection est induite sur l'annotation et le calcul des spécificités est appliqué aux catégories de l'annotation visée.

---

<sup>2</sup> Pour plus d'information sur *Le Trameur* : <<http://www.tal.univ-paris3.fr/trameur>>

<sup>3</sup> Site du projet *The Penn Treebank* : <<http://www.cis.upenn.edu/~treebank/>>

<sup>4</sup> Site du corpus national russe (*Национальный корпус русского языка*) : <<http://www.ruscorpora.ru/>>

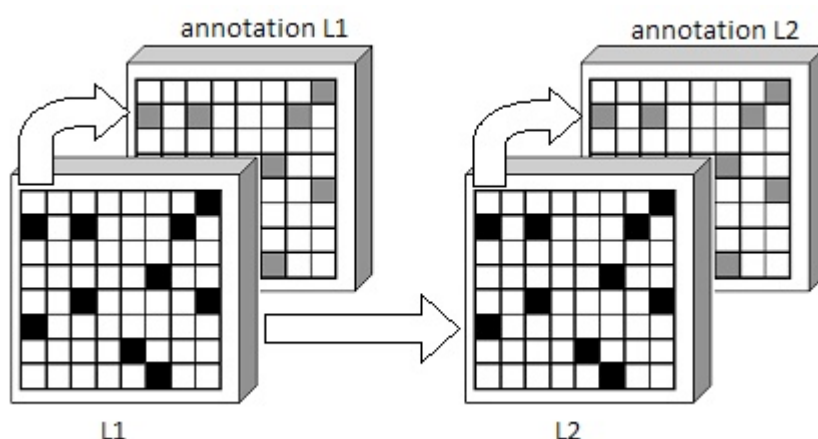


Figure 1. Approche systémique de la résonance textuelle

*Etape 1 : la mise en correspondance des textes originaux en anglais avec leurs traductions-adaptations en russe (un carré = texte ou annotation visé(e)).*

*Etape 2 : l'amorçage de la résonance au niveau lexical (L1 – induction – L2).*

*Etape 3 : la propagation par résonance vers le niveau d'annotation visé (exemple : annotation grammaticale) : L1 – induction – annotation L1 ; L2 – induction – annotation L2.*

Ainsi, pour les termes « people » | « человек », la sélection induite sur l'annotation morphosyntaxique de chaque volet du corpus a permis le recensement des figements caractéristiques de ces fragments. Après le filtrage des suites de catégories grammaticales qui se contiennent (exemple : la suite A+B+C, contenue dans A+B+C+D), les emplois suivants sont classés premiers dans chaque volet par ordre de l'indice de spécificité (cf. figure 2).

Selon la formalisation textométrique, il s'agit de *Trames* induites par résonance. Elles sont matérialisées par des retours aux contextes pour l'interprétation des structures lexico-grammaticales sous-jacentes :

	Élément caractéristique	Ind-Specif	Fq-Totale	Fq-Partie
Anglais	CD (nombre cardinal) + NNS (nom au pluriel) + VBD (verbe au passé) + VVN (participe passé)	30.1	326	269
Russe	S (nom) + V (verbe) + NUM (nombre) + S (nom)	21.4	476	215

(1)	2002-06-17.1 general-JJ on-IN friday-NP in-IN which-WDT 12-CD <b>people-NNS</b> were-VBD killed-VVN and-CC 2002-06-17.1 взрывчаткой-S машина-S . в-PR результате-S теракта-S погибли-V 12-NUM <b>человек-S</b> , 44-NUM 2002-06-17.1 подозрению-S в-PR организации-S атаки-S . показания-S дали-V трое-NUM свидетелей-S . сразу-ADV
(2)	2001-06-22.3 or-CC opposition-NN to-TO parades-NNS ." three-CD <b>people-NNS</b> were-VBD arrested-VVN for-IN 2001-06-22.3 мере-S 24-NUM полицейских-A пострадали-V . полиция-S задержала-V троих-NUM <b>человек-S</b> и-CONJ ,
(3)	2001-07-25.2 of-IN a-DT sect-NN office-NN . seven-CD <b>people-NNS</b> were-VBD killed-VVN and-CC 2001-07-25.2 метро-S . в-PR результате-S этих-A-PRO терактов-S погибли-V 19-NUM <b>человек-S</b> (семь-NUM
(4)	2003-07-02.1 monsoon-NN rains-NNS . in-IN may-MD , 38-CD <b>people-NNS</b> were-VBD killed-VVN when-WRB 2003-07-02.1 punjab-NP officials-NNS say-VVP at-IN least-JJS <b>six-CD of-IN the-DT dead-JJ were-VBD</b> 2003-07-02.1 при-PR этом-A-PRO смертельные-A ранения-S получили-V шесть-NUM <b>человек-S</b> , находящихся-V

Figure 2. Trames induites par résonance dans BBC\_Lenta.RU

On remarque que les *Trames* CD+NNS+VBD+VVN (en anglais) et S+V+NUM+S (en russe) sont sémantiquement liées dans les deux volets du corpus (cf. figure 2). Il y a mise en équivalence lorsqu'elles intègrent les termes « people » | « человек » qui se correspondent (cf. figure 2, exemples 1-2). Il s'agit de l'équivalence complète de schémas lexico-grammaticaux.

L'appariement est délicat en cas de non-correspondance entre « people » et « человек » (cf. figure 2, exemple 4). Dans ce cas, les indices récoltés au niveau morphosyntaxique ne sont pas assez précis pour l'appariement.

On relève aussi la problématique soulevée par des cas d'adaptation et de compression de l'information à traduire, le texte adapté en russe étant généralement plus sommaire que le texte original (cf. figure 2, exemple 3).

### 5.3. Croisement des annotations sur les Trames parallèles

Dans cette expérimentation, deux niveaux d'annotations sont croisés sur la *Trame*. Les sections qui suivent présentent les détails de la procédure de traitement appliquée au corpus *BBC\_Lenta.RU*.

#### 5.3.1. Induction itérative

Une sélection est amorcée dans le volet anglais du corpus *BBC\_Lenta.RU* à partir du terme « researchers » (91 occ.). On induit par résonance les formes « учёные » (133 occ.) et « учёных » (50 occ.) dans le volet russe. Ces termes sont en correspondance partielle. L'induction dans les fragments où « researchers » et « учёны(е|x) » ne se correspondent pas amène la découverte d'un réseau de correspondances traductionnelles multiples qui reflètent les phénomènes de synonymie : « researchers » et « scientists » en anglais, traduits par « учёны(е|x) » et « исследовател(и|ей) » en russe.

#### 5.3.2. Induction sur annotation

Pour amplifier le phénomène de résonance autour des pôles « researchers | scientists » et « учёные | исследователи », la forme graphique de chaque item concerné est mise en relation avec la catégorie morphosyntaxique associée. Le calcul des répétitions segmentales au niveau de l'annotation morphosyntaxique tient compte de ces nouvelles catégories hybrides créées pour les termes étudiés.

L'induction sur l'annotation morphosyntaxique révèle ensuite la présence caractéristique des suites des catégories suivantes :

pôle « researchers | scientists » + *VVP* (verbe au présent, différent de la 3ème personne sing.) [F = 72]  
 pôle « учёные | исследователи » + *V* (verbe) [F = 98]

#### 5.3.3. Extractions avec annotations croisées

Les patrons caractéristiques sont donnés en entrée au logiciel *Le Trameur* pour l'extraction des séquences textuelles qui y sont associées dans le corpus. L'extraction est menée en croisant deux niveaux d'annotation (formes et catégories morphosyntaxiques). En sortie, l'annotation retenue (formes) permet de représenter les séquences effectivement attestées dans le corpus (cf. figures 3-4).

Sur la figure 3, la visualisation graphique est réalisée à l'aide du programme *Pajek* (Batagelj, Mrvar, 2003). Il est appelé par *Le Trameur* après la génération des *Trames* associées aux patrons visés dans les deux langues, avec croisements d'annotations.

En analysant les graphes présentés sur la figure 3, on constate une plus grande variabilité des séquences associées au patron russe. Comparé à l'étiquetage de l'anglais, la catégorisation des verbes en russe est moins détaillée. Cependant, l'écart constaté est d'avantage lié à des transformations lexico-grammaticales qui s'opèrent lors de l'adaptation des actualités scientifiques en russe.

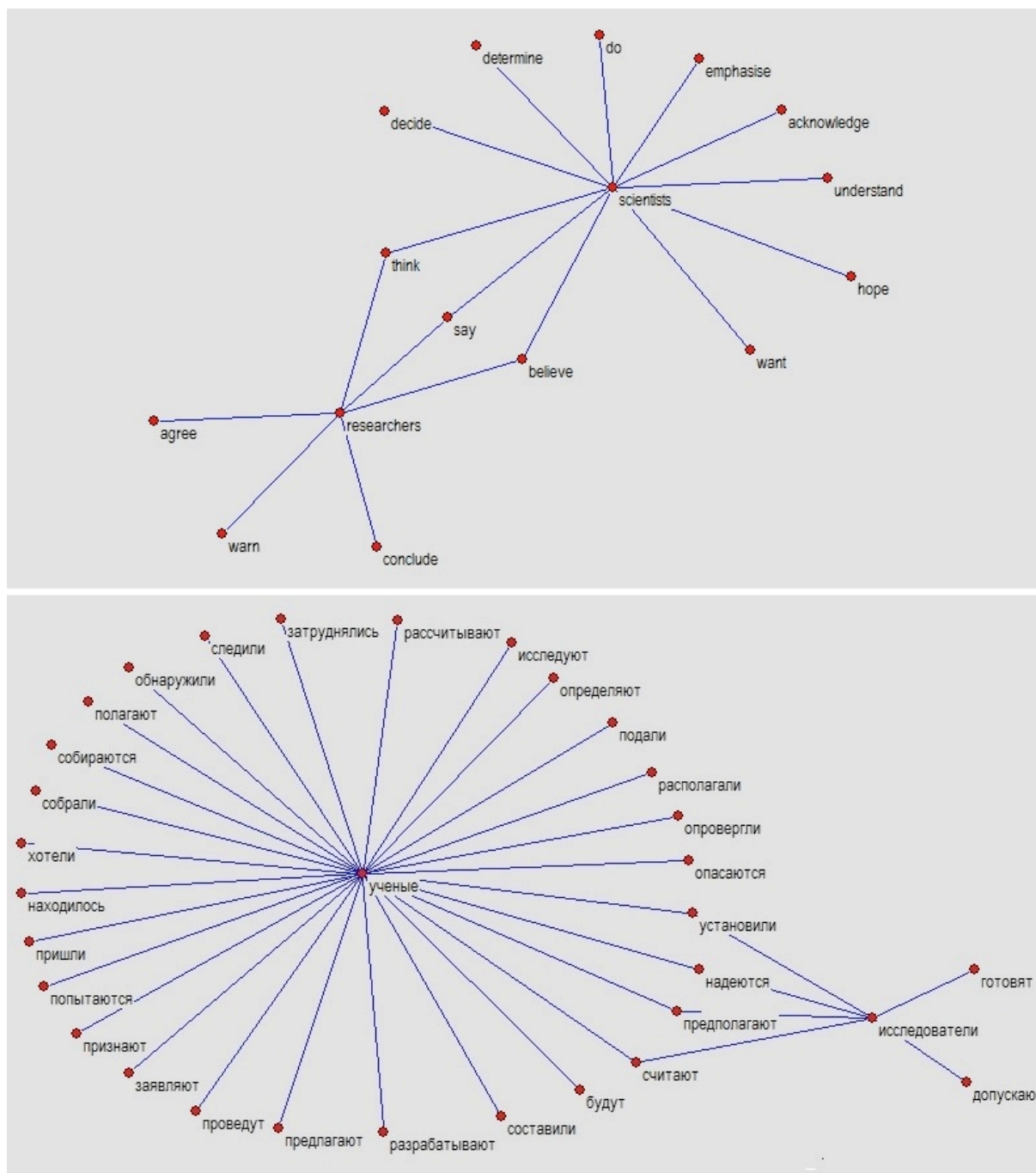


Figure 3. Trames « researchers | scientists » + VVP (ang.) et « учёные | исследователи » + V (rus.)

Sur la figure 4, la mise en correspondance des *Trames* représentées sur la figure 3, amène la découverte de contextes qui se correspondent dans les deux volets du corpus *BBC\_Lenta.RU* :

```

-----PARTIE{texte=2001-02-07.1}-----
scientists believe they have found a new mammal species a camel that lives in a
ученые обнаружили в тибете новую разновидность верблюдов, сообщает bbc

-----PARTIE{texte=2002-05-27.1}-----
reporter uk scientists say they could win the race to find proof of life on mars following
британские ученые полагают, что смогут первыми доказать существование жизни на марсе,

-----PARTIE{texte=2002-07-11.2}-----
poitiers, france. scientists say it is the most important discovery in the search for
международная группа учёных обнаружила в пустыне чада череп, который назван наиболее

-----PARTIE{texte=2002-09-07.2}-----
light-years away. scientists hope this will enable them to test one of einstein's assumptions
в субботу ученые проведут эксперимент, призванный доказать или опровергнуть общую теорию

-----PARTIE{texte=2003-12-29.1}-----
to communicate with beagle and scientists think this is their best hope of raising the robot.
сценарий исключительно маловероятен, и ученые будут продолжать пытаться "оживить" аппарат,

-----PARTIE{texte=2004-02-27.2}-----
some researchers think comet impacts may even have seeded the early earth with the chemistry
некоторые исследователи надеются найти на ядре подтверждение теории о космическом

```

Figure 4. Contextes « researchers | scientists » + VVP et « учёные | исследователи » + V

La visualisation des *Trames* à l'aide de la topographie textuelle ouvre des perspectives de navigation parallèle dans les corpus comparables avec exploration itérative de structures parallèles et prise en compte des écarts éventuels.

## 6. Topographie des *Trames* parallèles

Une représentation simultanée à la surface textuelle d'un ensemble d'éléments induits par résonance permet d'adopter un point de vue systémique dans l'analyse du bi-texte.

Sur les figures 5-6, l'induction est amorcée par le terme anglais « war » (1 462 occ.). Attestée dans 587 textes (sur 2 345 au total), il se situe parmi les dix mots pleins les plus fréquents du volet anglais du corpus *BBC\_Lenta.RU*. Le terme induit dans le volet russe est « войны » (224 occ.), c'est la forme fléchée de « война » (46 occ.).

Les *Trames* parallèles sont matérialisées à la surface du bi-texte à l'aide d'un marquage de couleur jaune effectué par *Le Trameur*. Chaque *Trame* est constituée de l'ensemble des figements morphosyntaxiques (répétitions segmentales des catégories) qui intègrent les termes de la sélection initiale (« war » | « войн(ы|а) »).

Sur chaque plan topographique, les suites de catégories se superposent en réalisant de nouveaux enchaînements de catégories, caractéristiques des fragments textuels sélectionnés. Ces enchaînements sont ensuite retranscrits au niveau des occurrences de formes graphiques, donnant lieu au repérage parallèle des éléments d'équivalences traductionnelles dans les deux volets du corpus *BBC\_Lenta.RU* (cf. figures 5-6).



Le Trameur - Le Mètre Lexicométrique @CLA2T-P3 V. 11.24

Shift-clic sur carré : affichage | clic-droit sur carré : spécificités | Control-clic sur carré : sélection | Shift-Control-clic sur sélection : désélection

Seuillage : 1 5 10 ++ | Modifier seuillage :

Cadre Ventilation Section Forme-Lemme Catégorie-Tag Segment Cocc Stat Concordance Patron Graphe Sélection Rapport Param

Mo L. Sections sélectionnées : 0 N° Sect. : 1998:(1832166,1832220) Annotation : 1 Aperçu : 50

4 in his weekly radio message  
 5 broadcast on stations across the us on saturday  
 6 mr bush said he would "reach out" to allies and sceptics at home and abroad.  
 7 he stressed the continuing importance of the war on terror and the struggle against disease and hunger and poverty.  
 8 on domestic issues he pledged tax reforms and a clampdown on lawsuits.  
 9 with more than two months until the official start of his second term at the presidential inauguration on 20 january  
 10 mr bush used his message to address republicans and democrats.  
 11 we have one country  
 12 one constitution  
 13 and one future that binds us. and when we come together and work together  
 14 there is no limit to the greatness of america.  
 15 us president george w bush  
 16 he steered clear of triumphalism  
 17 saying that supporters of both parties could agree a common approach to the war on terror.  
 18 "americans are expecting bipartisan effort and results  
 19 the president said.  
 20  
 21 whatever our past disagreements  
 22 we share a common enemy and common duties."  
 23 "every civilised country has a stake in the outcome of this war  
 24 mr bush added, insisting he would  
 25 continue reaching out" to nato and european nations whose relations with the us have been strained by the war in iraq.  
 26 'shared responsibilities'  
 27 on the home front  
 28 mr bush described the challenge of the coming four years in reserved tones  
 29 speaking of "serious responsibilities and historic opportunities".  
 30 on domestic policy there were no hints of any plans to use a second term to adopt a more conservative social agenda.  
 31 instead mr bush repeated the message of his victory speech on wednesday: "to make this nation stronger and better  
 32 i will need the support of republicans and democrats and independents  
 33 and i will work to earn it."  
 34 mr bush focused on what he termed "frivolous" lawsuits and an "outdated" tax code - two issues he repeatedly stressed while on the campaign  
 35 trail.

Position: <1832511>  
 Forme: <war> | Freq: 1462  
 Lemme: <war> | Freq: 1544  
 Cat: <NN\_war> | Freq: 1462

Shift-Clic : sélection | Clic-droit : édition | Ctrl-Clic : newed | 2-Clic : graphe | Shift-Clic-droit : relation | Control-Clic-droit : recherche relation

Figure 5. Indices récoltés pour l'appariement dans le volet anglais BBC\_Lenta.RU

Le Trameur - Le Métier Lexcométrique @CLA21-P3 V. 11.24

Shift-clic sur carré : affichage | clic-droit sur carré : spécifiés | Control-clic sur carré : sélection | Shift-Control-clic sur sélection : désélection

Seuillage : 1 5 10 ++ | Modifier seuillage :

Control-clic sur marqueur de page : sélection 3 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Sections sélectionnées : 0 N° Sect. : 1988:(719)26,719716 Annotation : 1 Aperçu : 50

1 -->vir.lenta.ru -->что буш грядущий нам готовит?04.11.2004перспективы российско-американских отношений в ближайшие четыре года  
 2 президент США Джордж Буш призвал республиканцев и демократов объединить усилия во внешней и внутренней политике, чтобы победить терроризм за  
 3 время его второго президентского срока. Как сообщает BBC News, в своем еженедельном радиовыступлении в субботу Буш подчеркнул, что он  
 4 обращается как к своим союзникам, так и к скептикам, в  
 5 Америке и за рубежом.  
 6 Он отметил важность продолжения войны с террором, а также борьбы с болезнями, голодом и бедностью. касаясь внутренних дел, Буш побещал  
 7 осуществить налоговую реформу и сократить поток судейских жалоб в некоторых сферах.  
 8 в своем выступлении президент обратился к обеим партиям - республиканцам и демократам. Он призвал их выработать единый подход к войне с  
 9 террором. "Каковы бы ни были наши разногласия в прошлом, у нас общий враг и общие обязанности", - подчеркнул Буш.  
 10 он добавил, что на исход этой войны может повлиять "каждая цивилизованная страна", и побещал, что он будет налаживать отношения с  
 партнерами по НАТО и еврозоне, с которыми у США существовали противоречия по поводу войны в Ираке.

Position: <719706>  
 Forme: <война> | Freq: 224  
 Lemme: <война> | Freq: 363  
 Cat: <S\_война> | Freq: 325

Figure 6. Indices récoltés pour l'appariement dans le volet russe BBC\_Lenta.RU

## 7. Conclusions et perspectives

Les résonances sur annotations ouvrent des perspectives de navigation dans les corpus comparables avec une prise en compte d'emplois parallèles à plusieurs niveaux d'analyse linguistique. Les premiers appariements réalisés au niveau lexical par résonance textuelle servent d'indices pour l'induction sur les annotations.

Les observations statistiques du bi-texte confirment les principes énoncés par la linguistique systémique fonctionnelle qui souligne que la morphologie, la syntaxe et le lexique correspondent au même niveau d'analyse : la *lexicogrammaire*. Les croisements des annotations qui suivent ces principes amènent la résonance textuelle amplifiée.

Formalisés sous forme de *Trames* textométriques parallèles, les figements morphosyntaxiques recensés automatiquement dans le voisinage contextuel des pôles lexicaux appariés matérialisent des schémas lexico-grammaticaux parallèles. Leur visualisation à la surface de textes est facilitée par la topographie textuelle.

Comme chaque bi-texte est un système d'équivalences de fonctions qui se réalise dans un contexte social donné, c'est l'approche fonctionnelle de l'alignement textuel qui peut contribuer à l'avancement de la réflexion sur l'informatisation des liens dans les corpus multilingues. Vu sous cet angle, un alignement textuel est multiple. Dans sa construction, il privilégie la mise en correspondance d'éléments significatifs pour l'équivalence de sens dans une situation de communication donnée. C'est donc en faisant émerger du corpus des traits linguistiques caractéristiques que l'on parvient à percer dans l'interprétation des axes parallèles, notamment dans les corpus à degrés de comparabilité variables.

De ce point de vue, les propositions récentes de description et d'implémentation des objets textométriques (Fleury, 2013) trouvent des applications concrètes pour la formalisation des phénomènes de résonance textuelle multilingue lors de la navigation en corpus.

## Références

- Banks D. (2005). *Introduction à la linguistique systémique fonctionnelle*. Editions L'Harmattan.
- Batagelj V. and Mrvar A. (2003). Pajek - Analysis and Visualization of Large Networks. In Jünger M. and Mutzel P., editors, *Graph Drawing Software*. Springer, Berlin, pp. 77-103.
- Fleury S. (2013). Le Trameur. Propositions de description et d'implémentation des objets textométriques. *Publication sur le site de l'Université Paris 3* <<http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definITIONS-objets-textometriques.pdf>>
- Fleury S. et Zimina M. (2008). Utilisation de MkAlign pour la traduction philologique. *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT'08)*, pp. 483-493.
- Gledhill C. (2011). The lexicogrammar approach to analysing phraseology and collocation in ESP texts. *ASP*, 59: 5-23.
- Klementiev A. et Roth D. (2006). Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 817-824 <[http://cogcomp.cs.illinois.edu/page/resource\\_view/1](http://cogcomp.cs.illinois.edu/page/resource_view/1)>
- Lebart L. et Salem A. (1994). *Statistique textuelle*. Dunod.
- Manfredi M. (2008). *Translating Text and Context: Translation Studies and Systemic Functional Linguistics*. Volume 1: Translation Theory. Quaderni del CeSLiC: *Functional Grammar Studies for Non-Native Speakers of English*. Bologna, DUPress.

- Halliday M. (1961). Categories of the theory of grammar. *Word*, 17(3): 241-292.
- Plassard F. (2007). *Lire pour traduire*. Presses Sorbonne Nouvelle.
- Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris, Klincksieck.
- Salem A. (2004). Introduction à la résonance textuelle. *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT'04)*, pp. 986-992.
- Salem A. et Fleury S., editors, (2009). *Lexicometrica*, numéro spécial « Explorations textométriques »  
<<http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm>>
- Schmid H. (1994). Probabilistic Part-of-speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, pp. 44-49.
- Tiedemann J. (2011). *Bitext Alignment*. Morgan and Claypool Publishers (USA).
- TransRead ANR-12-CORD-0015 (en cours) <<http://transread.limsi.fr/>>
- Zimina M. (2004). L'alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles. *Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles (JADT'04)*, pp. 1195-1202.
- Zimina M. (2005). Bi-text Topography and Quantitative Approaches of Parallel Text Processing. *Proceedings from the Corpus Linguistics Conference Series*, 1(1).