

BI-TEXT: A SYSTEMIC FUNCTIONAL APPROACH AND TEXTOMETRIC ANALYSIS

MARIA ZIMINA mzimina@eila.univ-paris-diderot.fr

PARIS DIDEROT, CLILLAC-ARP

PLAN

- What is *bi-text*? [Harris, 1988]
- How to align multilingual texts?
- The ‘Lexicogrammar’ approach to aligning comparable English/Russian corpora (*BBC_LENTA.RU*)
- Principles of textometric analysis
 - Text segmentation and annotations
 - Incremental textual resources
 - Textual resonance
 - Bi-text topography

WHAT IS BI-TEXT?

"What is *bi-text*? [...] I offer it to translators as a new concept in translation theory.[...]

One way to describe bi-text, therefore — and this is a basic definition — is to say that it is ST and TT as they co-exist in the translator's mind at the moment of translating. [...]

Another way of putting it is to say that a bi-text is not two texts but a single text in two dimensions, each of which is a language. [...]

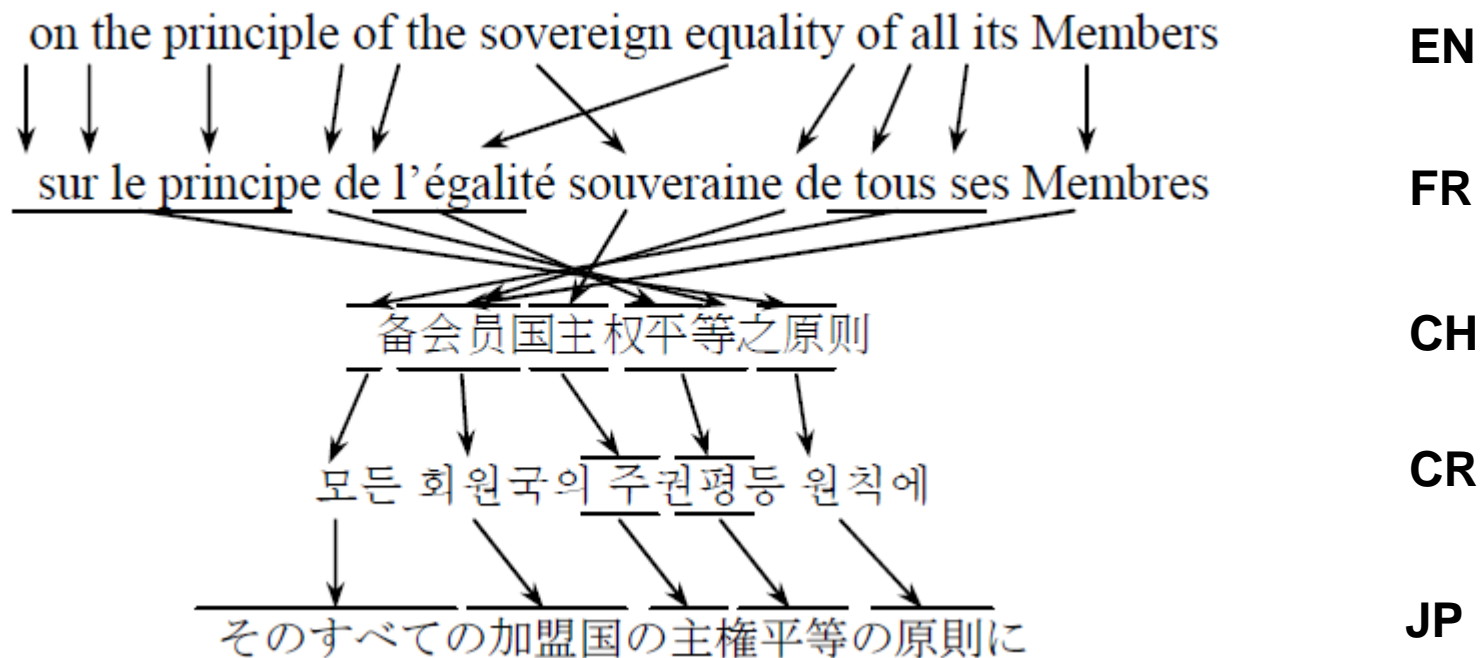
To semioticians, by the way, I submit that *bi-text* falls into the same paradigm as *intertext*, in that it is a construct of two or more related texts."

Bi-text, a new concept in translation theory.

Brian Harris, 1988

HOW TO ALIGN MULTILINGUAL TEXTS?

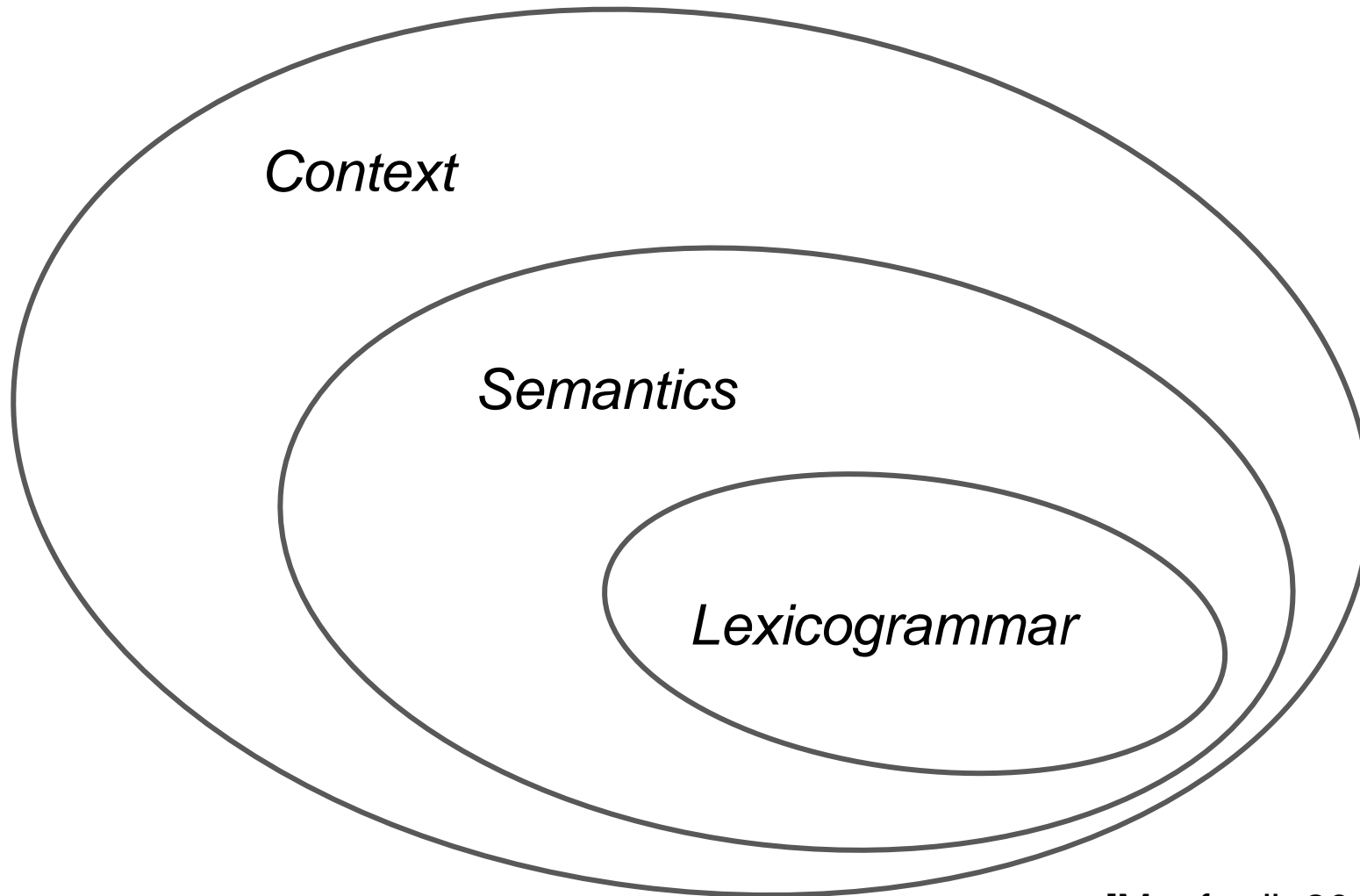
“configurations of multidimensional meanings, rather than [...] containers of *content*” [Steiner and Yallop, 2001]



[CHO, 2009]

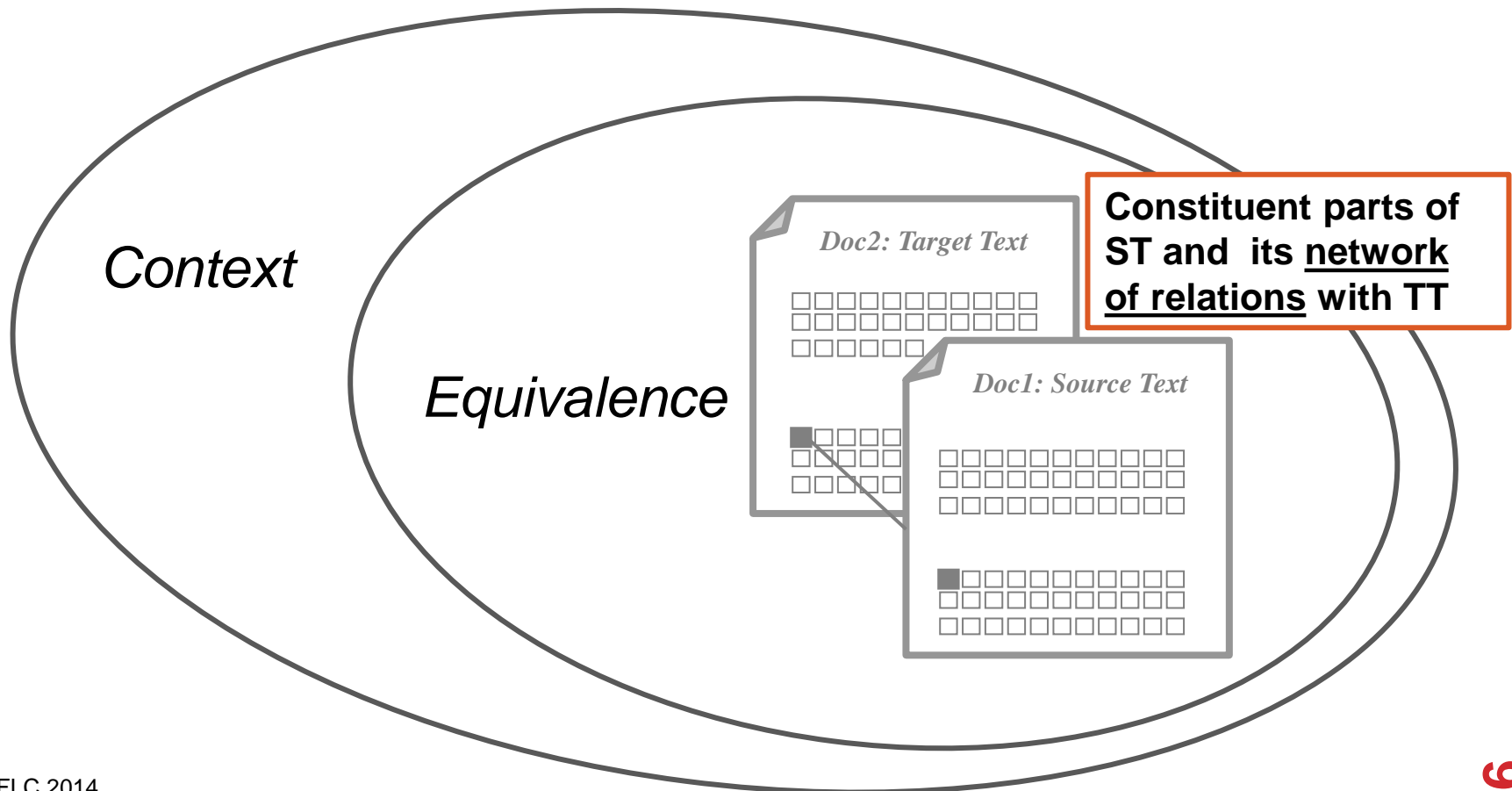
Lexicometrica, vol. 3 “Corpus multilingues”

SFL: ANALYTICAL TOOL FOR BI-TEXT ANALYSIS



BI-TEXT NETWORK

“If meaning is function in context, [...] then equivalence of meaning is equivalence of function in context” [Halliday, 1992].



COMPARABLE CORPUS

BBC-LENTA_RU (2 345 TEXTS)

BBC News Feeds (2001-2005) and their
adapted translations from *Lenta.ru*

English	Russian
1 million words	500 000 words

[Klementiev and Roth, 2006]

[http://cogcomp.cs.illinois.edu/page/resource_view/1]

TEXT ALIGNMENT AND LEXICOGRAMMAR

Extended **lexicogrammatical patterns** as a key feature in text alignment:

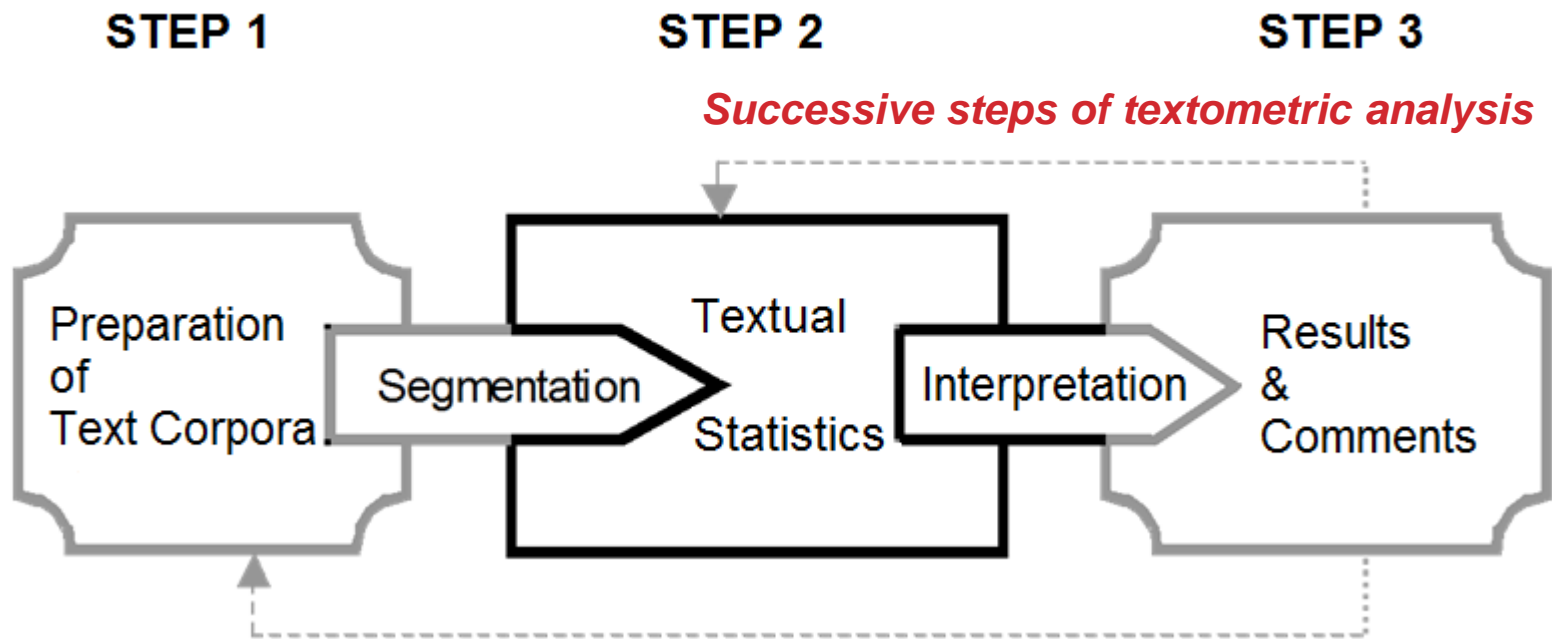
English	Russian
<p><texte = 2004-11-07.0> us president george w bush has called for joint efforts at home and abroad to achieve his second term goals and win the war on terror. mr bush said he would "reach out" to allies and sceptics at home and abroad. /.../</p>	<p><texte = 2004-11-07.0> президент сша джордж буш призвал республиканцев и демократов объединить усилия во внешней и внутренней политике, чтобы победить терроризм за время его второго президентского срока. как сообщает bbc news, в своем еженедельном радиовыступлении в субботу буш подчеркнул, что он обращается как к своим союзникам, так и к скептикам, в америке и за рубежом. /.../</p>

TEXTOMETRIC ANALYSIS (1/2)

Object of study:

- Text (or texts)
- **Objective:**
 - Define and *count* units in texts
- **Means:**
 - Tools, methods of **Textual Statistics**

TEXTOMETRIC ANALYSIS (2/2)

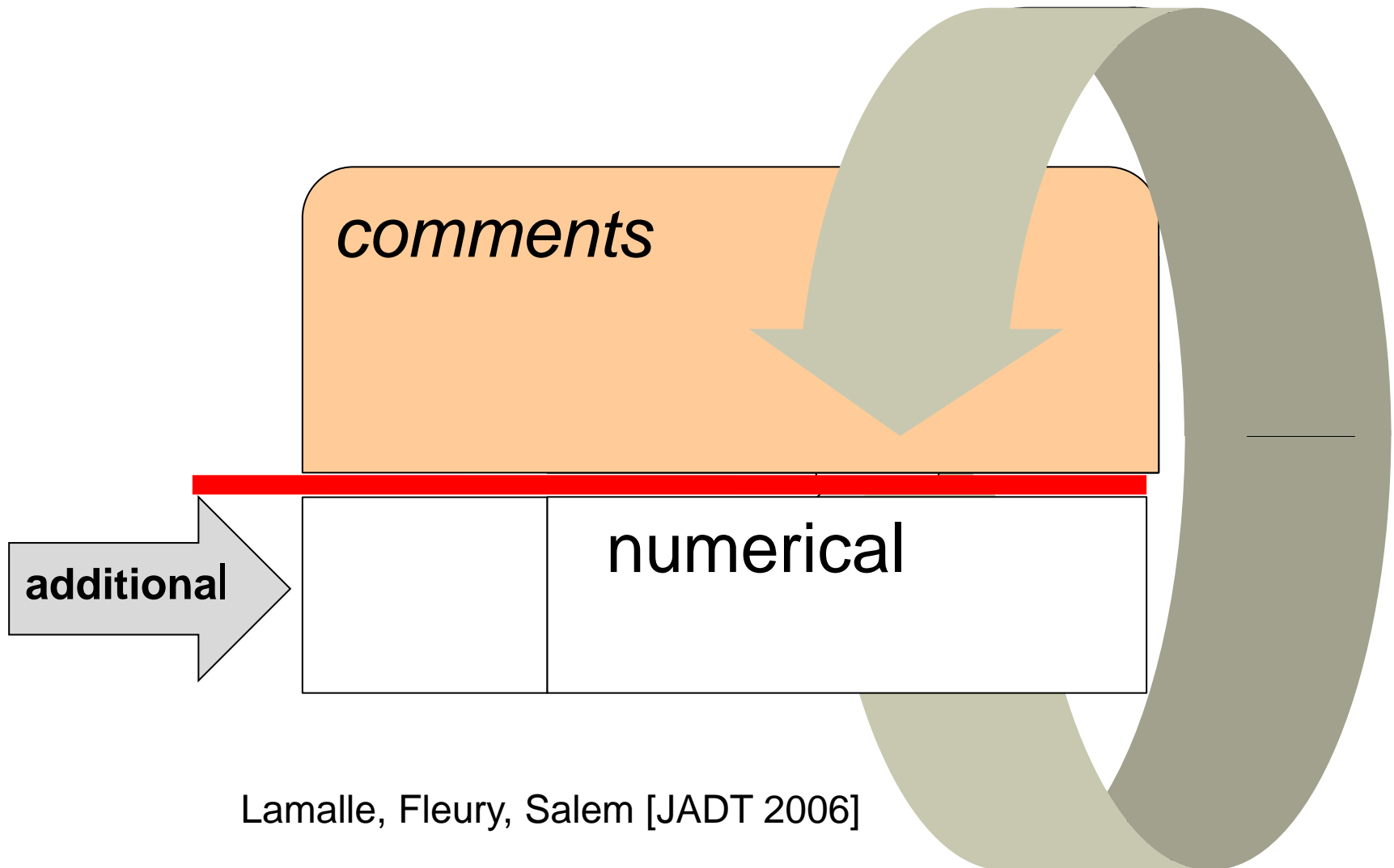


Lamalle, Fleury, Salem [JADT 2006]

TEXTOMETRIC PROCEDURES

- **Type 1** procedures:
 - *Fully automated*
- **Type 2** procedures:
 - *Require human intervention*
 - *Rely on complex textual resources*

INCREMENTAL TEXTUAL RESOURCES



Lamalle, Fleury, Salem [JADT 2006]

TEXT FLOW

Thread (in French: *trame*): sequence of annotated items with position identifiers

Annotation importée n°1	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)	
Catégorie		Cat(Le)		Cat(dormeur)		Cat(du)		Cat(val)	...
Lemme		Lemme(Le)		Lemme(dormeur)		Lemme(du)		Lemme(val)	...
Forme		Le		dormeur		du		val	...
Positions	1	2	3	4	5	6	7	8	...

Frame : selected spans of text defined on the *Thread*.

<STRUCTURE="TITRE"> Le dormeur du val
 <STRUCTURE="TEXTE POEME">
 <LIGNE="VERS1"> C' est un trou de verdure où chante une rivière,
 <LIGNE="VERS2"> Accrochant follement aux herbes des haillons
 <LIGNE="VERS3"> D' argent où le soleil ; de la montagne fière,
 <LIGNE="VERS10"> Sourirait un enfant malade, il fait un somme :
 <LIGNE="VERS11"> Nature, berce-le chaudement : il a froid.
 <LIGNE="VERS12"> Les parfums ne font pas frissonner sa narine ;
 <LIGNE="VERS13"> Il dort dans le soleil, la main sur sa poitrine,
 <LIGNE="VERS14"> Tranquille. Il a deux trous rouges au côté droit.
 <STRUCTURE="AUTEUR"> Arthur Rimbaud

TEXT ANNOTATIONS

Lemmatisation, morpho-syntactic tagging

- Integrated into *Trameur* via *TreeTagger* [Schmid, 1994]

Semantic annotation, etc.

- Possibility to implement multi-level annotations on the *Thread*

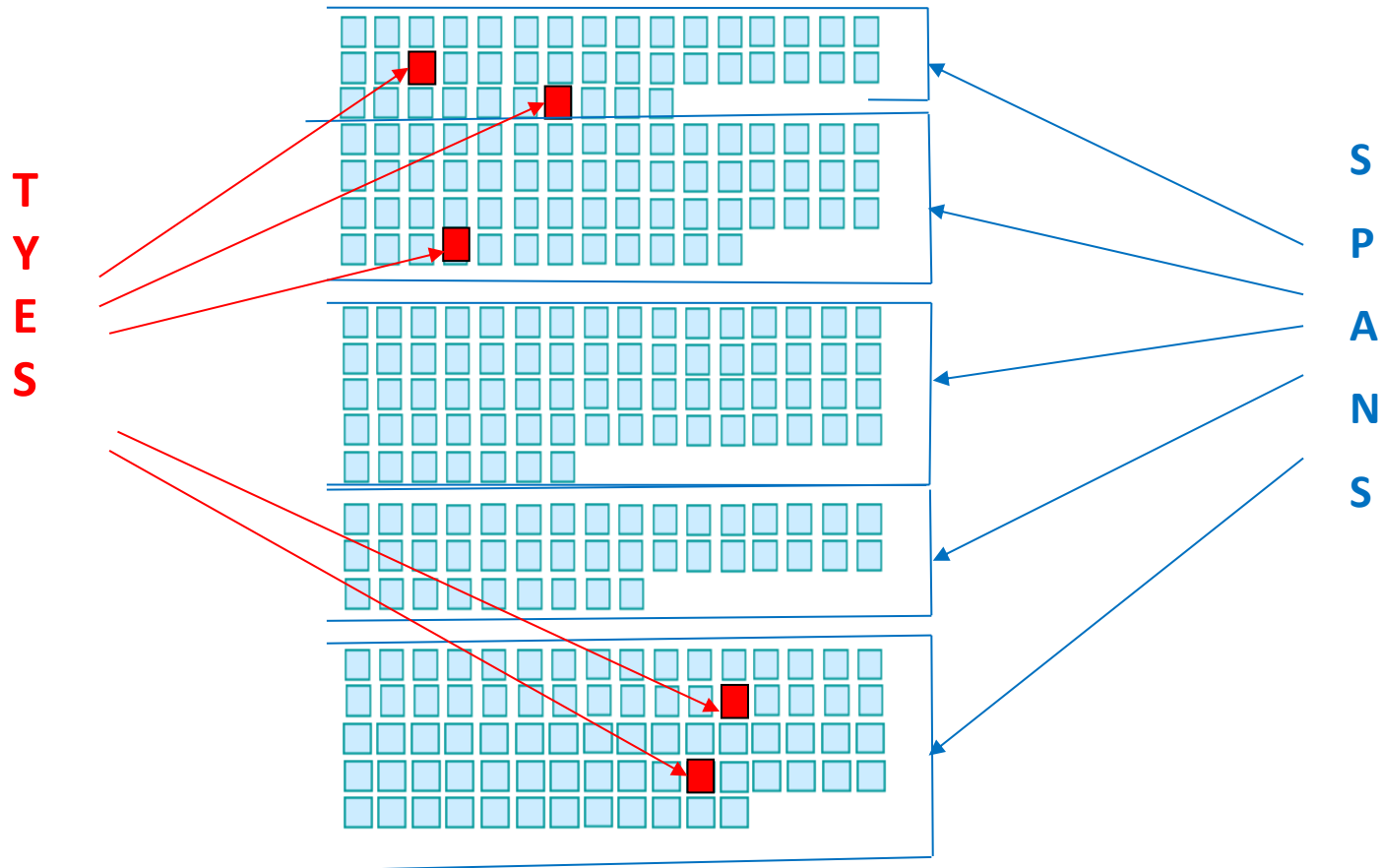
Annotation of dependency relations

***Thread* = sequence of annotated items with position identifiers**

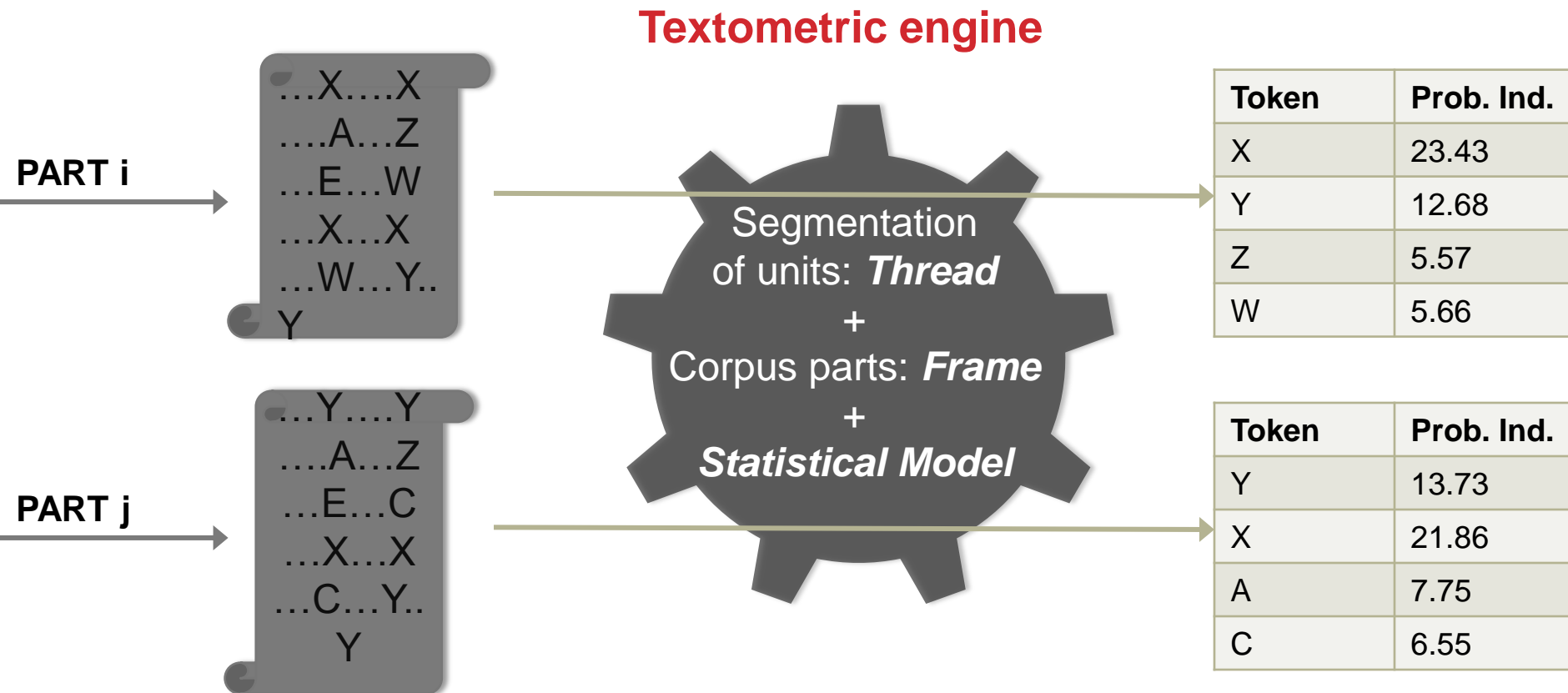
Son histoire a connu une fin tellement merveilleuse qu' elle mérite d' être contée dans le temps des Fêtes .

```
Position: <40>
Forme: <histoire> | Freq: 6
Lemme: <histoire> | Freq: 10
Cat: <NOM> | Freq: 3801
a-4: <21> | Freq: 1
a-5: <-> | Freq: 36858
a-6: <Topical_entity(23)> | Freq: 1
a-7: <SUJ(23)> | Freq: 1
```

STATISTICAL TABLES TYPES & SPANS



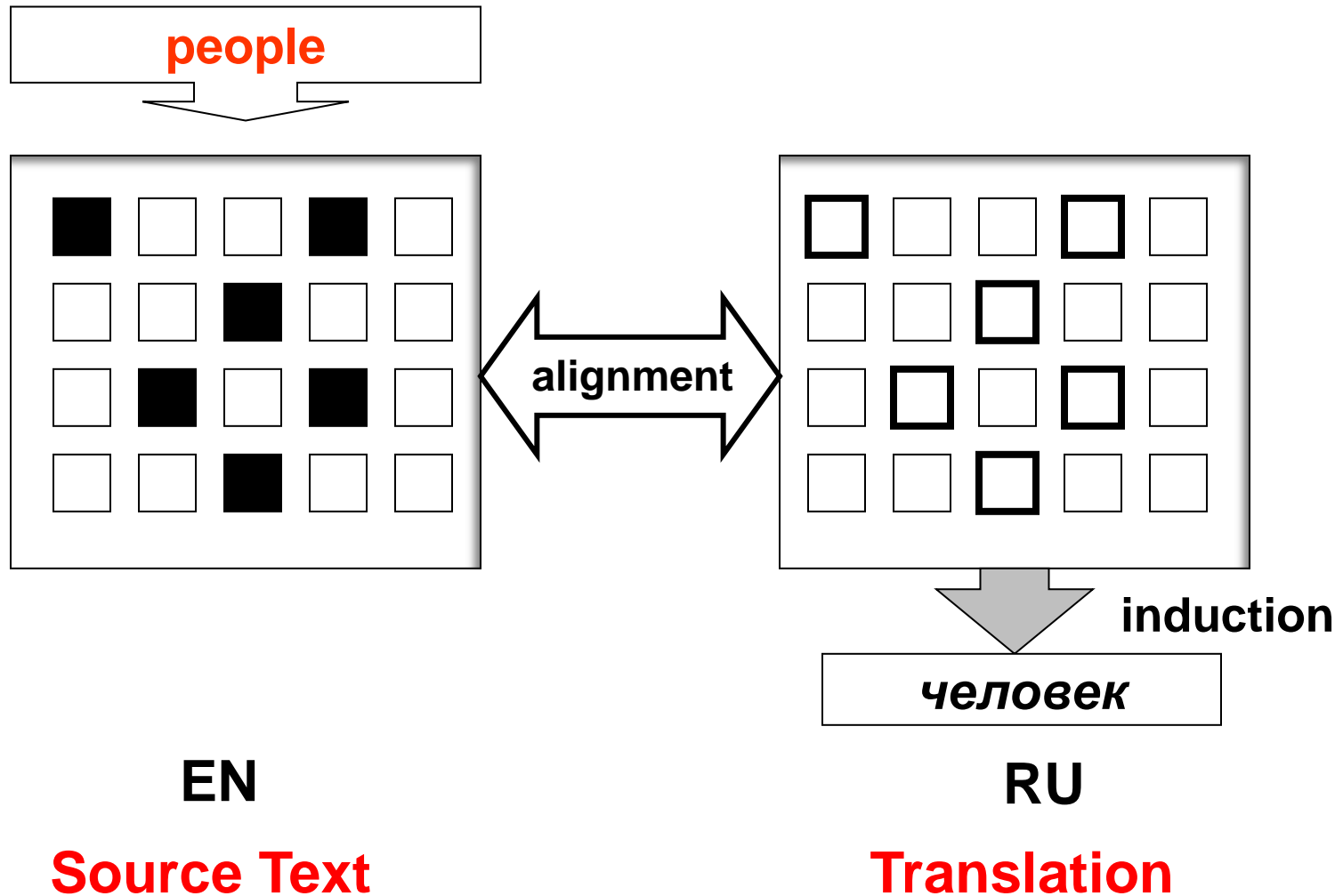
TRAMEUR [S. FLEURY, 2013]



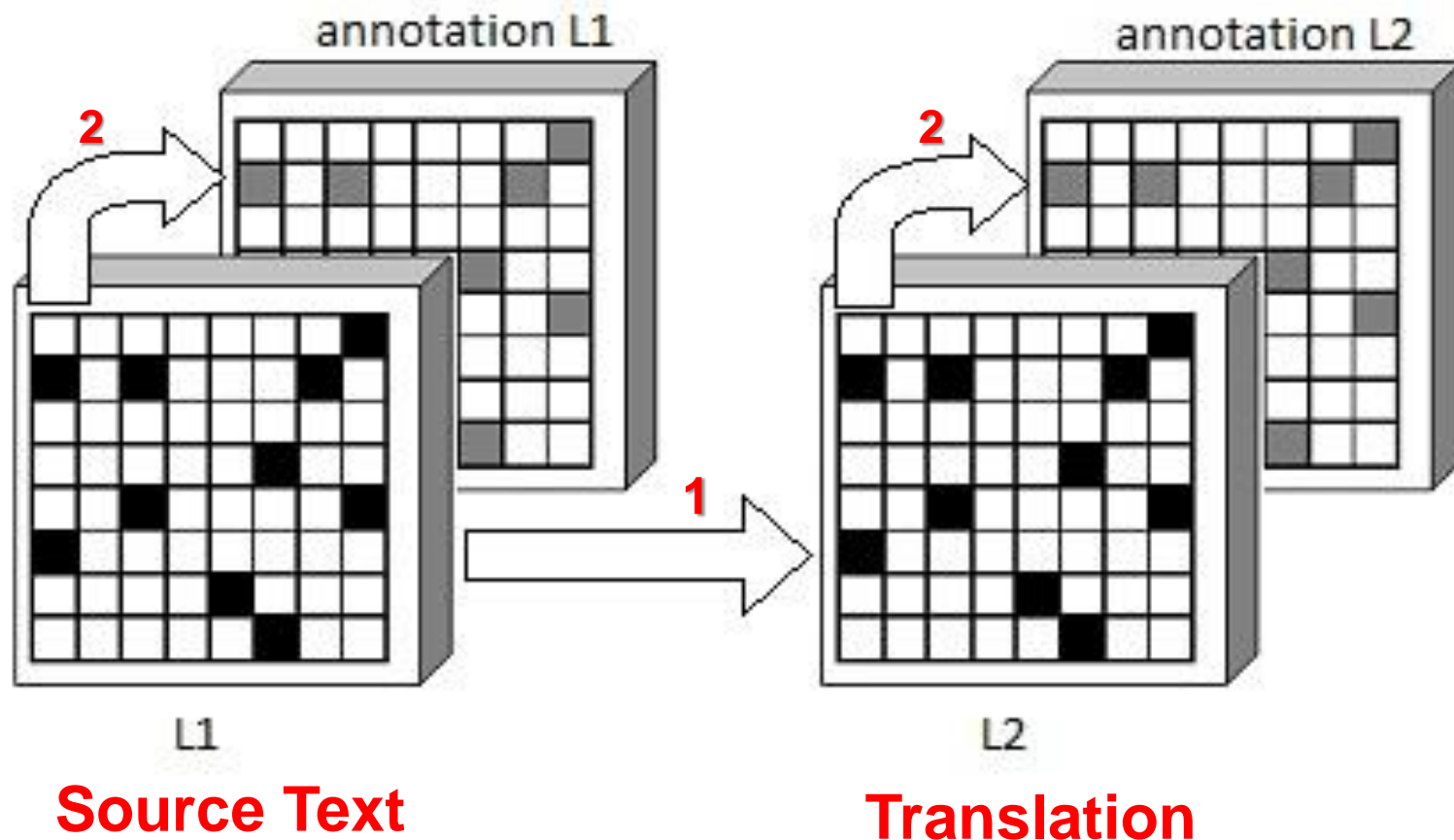
<http://www.tal.univ-paris3.fr/trameur>

TEXTUAL RESONANCE

Salem, Zimina [JADT 2004]



RESONANCE PROPAGATION TO ANNOTATIONS



CROSS-ANNOTATIONS AND REPEATED SEGMENTS

The "Lexicogrammar" approach

Repeated segment	Prob. Ind.	Repeated segment	Prob. Ind.
CD people VBD VVN	43.2	S V NUM человек	46.4
people VBP	42.9	ADV NUM человек	38.6
people WP	41.2	человек V PR	32.6
people IN DT	36.1	S V NUM	28.9
CD people IN	35.7	V PR NUM человек	28.5
CD people VHP	35.7	PR NUM S человек	27.9
IN JJS CD people	35.4	человек V V	24.9
JJS CD people	35.4	NUM человек V PR	24.9

NGram filter (**ABC**, **ABCD** ...) [Salem, 1987]

EXPLORING MULTI-LEVEL ANNOTATIONS

Choix des items du patron

Recherche croisée de patrons d'annotations

Insérer les noms des constituants du patron
et le niveau d'annotation visé :

Pour chaque item du patron :
Insérez la valeur de l'item et le niveau de l'annotation associée

ITEM(1)	CD	3
ITEM(2)	people	1
ITEM(3)	VBD	3
ITEM(4)	VVN	3
Annotation en sortie.....	Forme	1

Annuler **BBC**

CD + **people** + VBD + VVN
(English)

S + V + NUM + **человек**
(Russian)

Choix des items du patron

Recherche croisée de patrons d'annotations

Insérer les noms des constituants du patron
et le niveau d'annotation visé :

Pour chaque item du patron :
Insérez la valeur de l'item et le niveau de l'annotation associée

ITEM(1)	S	3
ITEM(2)	V	3
ITEM(3)	NUM	3
ITEM(4)	человек	1
Annotation en sortie.....	Forme	1

Annuler **LENТА_RU** Enregistrer

TRAMEUR

ALIGNMENT AND LEXICOGRAMMAR

CD + **people** + VBD + VVN (English)

S + V + NUM + **человек** (Russian)

-----BBC=2001-07-25.2-----
of a sect office. **seven people were killed** and

-----LENTA_RU=2001-07-25.2-----
в результате этих терактов погибли **19 человек** (семь в мацумото

-----BBC=2001-06-22.3-----
or opposition to parades." **three people were arrested** for

-----LENTA_RU=2001-06-22.3-----
24 полицейских пострадали. полиция задержала троих **человек**

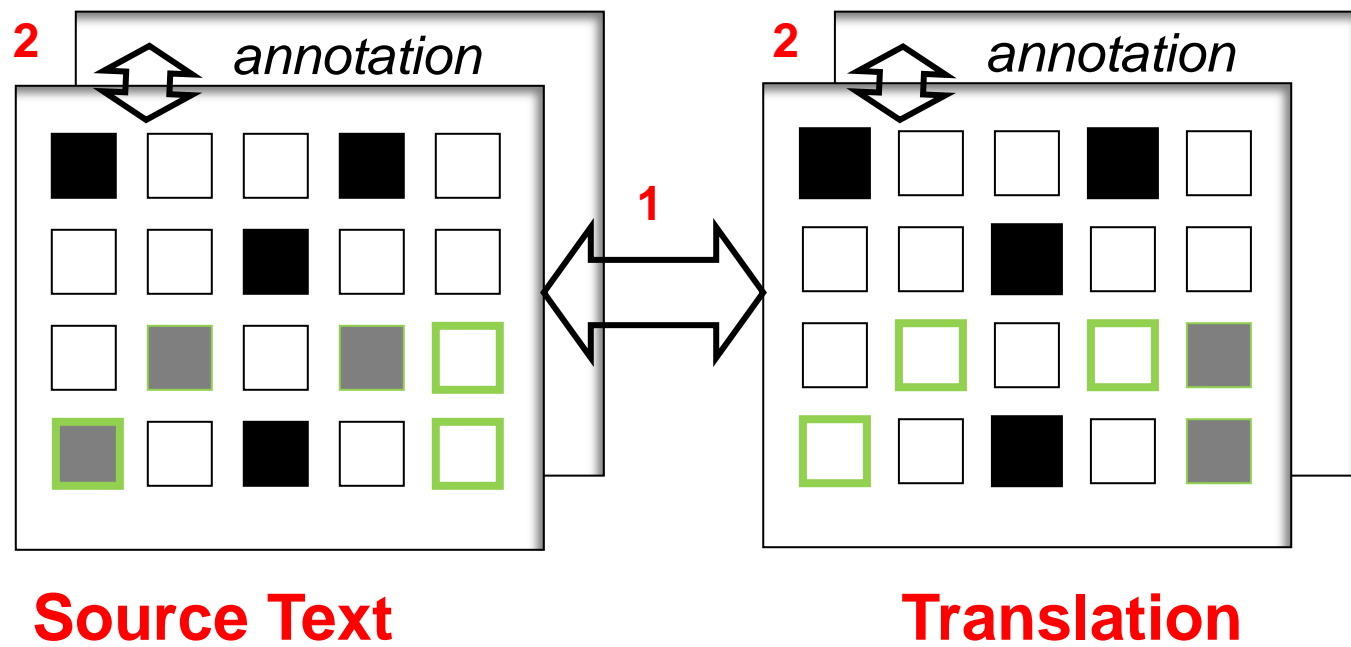
-----BBC=2002-06-17.1-----
general on friday in which **12 people were killed** and

-----LENTA_RU=2002-06-17.1-----
взрывчаткой машина. в результате теракта погибли **12 человек**

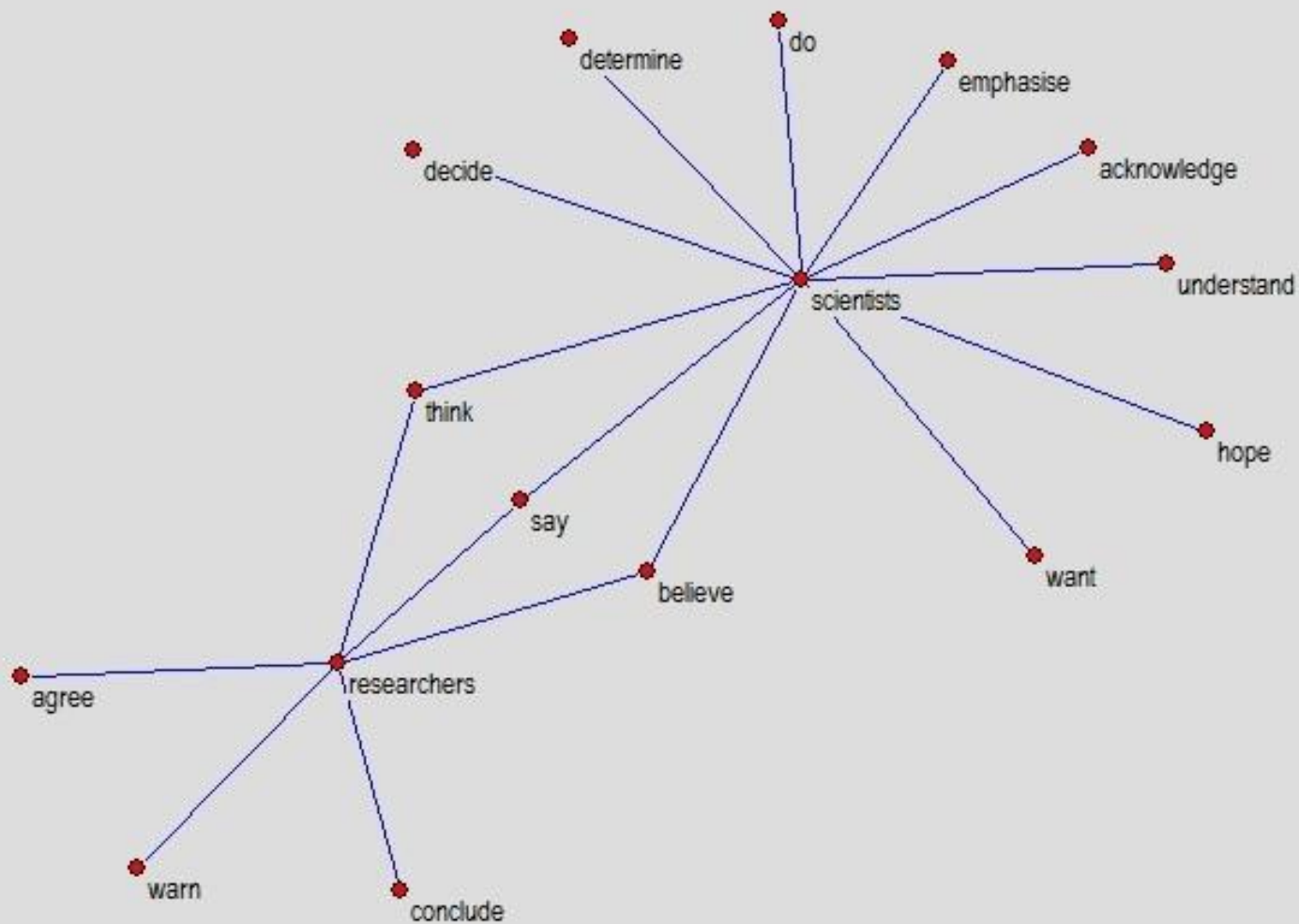
ITERATIVE INDUCTION

Contextually conditioned variants

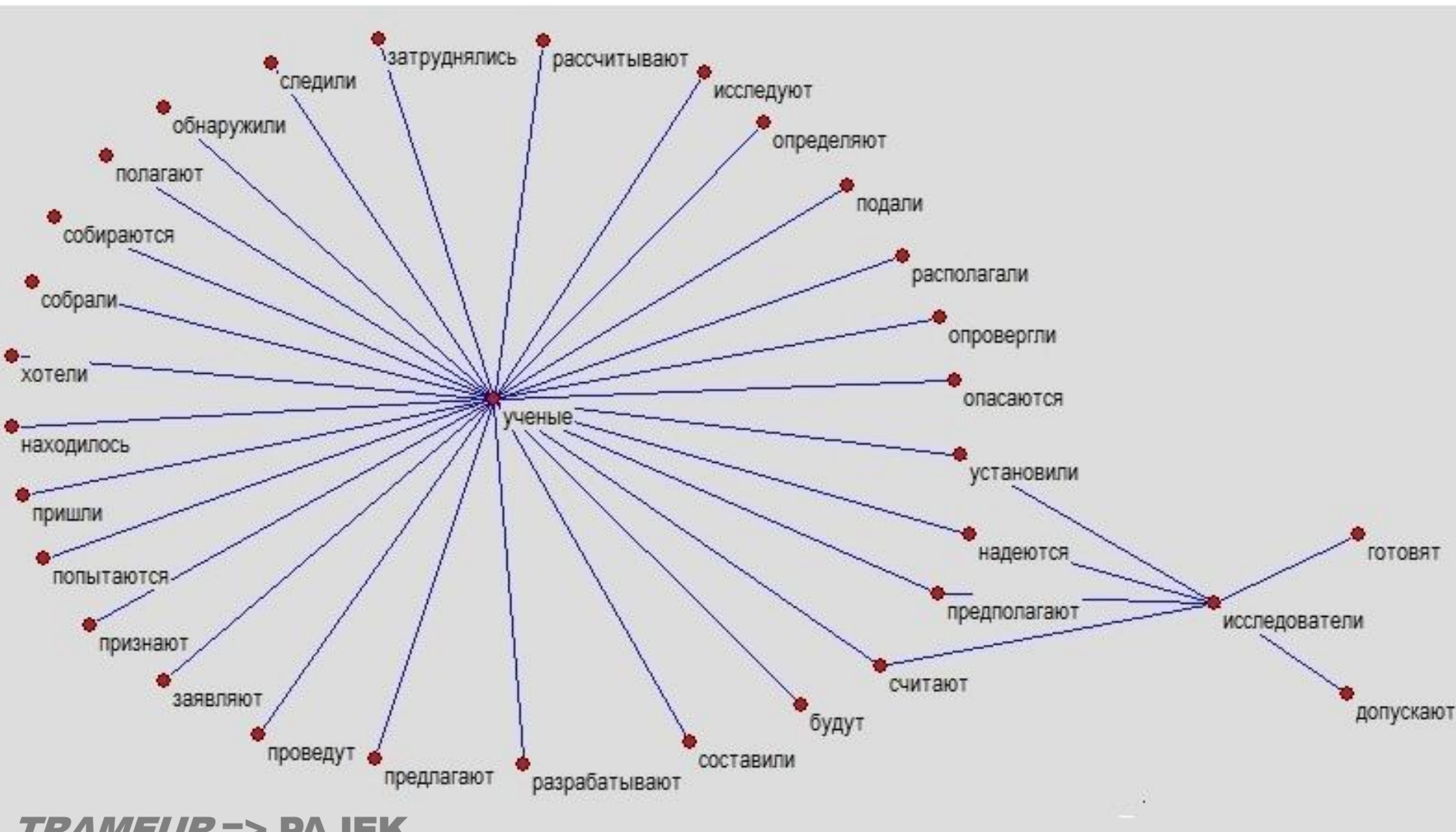
researchers \longrightarrow учёны(*e|x*)
scientists \longleftarrow \longrightarrow *исследователь*(*u|eй*)



[Zimina, 2004]



TRAMEUR => PAJEK



RESEARCHERS / SCIENTISTS + VVP

УЧЁНЫЕ / ИССЛЕДОВАТЕЛИ + V

-----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

TEXTOMETRIC INVENTORIES AND LEXICOGRAMMAR

NOUN + PREPOSITION + DETERMINER + WAR (English)

Le Trameur - Le Métier Lexicométrique @CLA2T-P3 V. 11.44

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

Concordance d'item :
[sélection po] RegExp ⓘ
(Ecrire motif supra puis <Entrée>)

Sélection Annotation :
 Forme Lemme Catégorie
n°Annotation
Annotation sélectionnée :
1 Forme
Fenêtre concordance :
10
Coloration annotation

Parties
texte

Export Concordance :

Ajouter au Rapport :

contexte-partie Tri-Concordance

Shift-Clc(pôle) : sélection | Clic-droit(pôle) : édition des annotations | Shift-Clc-droit(pôle) : relation | Ctrl-Clc-droit(pôle) : recherche relation Aperçu

-----PARTIE{texte=2003-07-07.2}-----
later in its first official **assessment** of the war the
-----PARTIE{texte=2003-07-08.2}-----
in the times suggested **public support** for the war in
-----PARTIE{texte=2003-07-13.0}-----
for apologising for japan's **role** in the war and
-----PARTIE{texte=2003-07-18.3}-----
mr blair also called **for vigilance** in the war against
-----PARTIE{texte=2003-07-29.1}-----
are almost invariably followed **by robbery** in this war. annan
-----PARTIE{texte=2003-07-29.2}-----
zvornik civilians killed **at the start** of the war said
-----PARTIE{texte=2003-07-30.0}-----
he did not believe **an inquiry** into the war on
-----PARTIE{texte=2003-08-01.0}-----
final approval to a **new version** of a war crimes
-----PARTIE{texte=2003-09-04.0}-----
bosch said dr kelly's **view** on the war was
-----PARTIE{texte=2003-12-20.2}-----
has not been **at the centre** of the war on
-----PARTIE{texte=2004-02-04.0}-----
organised the protest because **of anger** over the war and
-----PARTIE{texte=2004-02-09.2}-----
to mr djindjic's co-**operation** with the war crimes
-----PARTIE{texte=2004-04-05.2}-----
south east asia its **second front** in the war on
-----PARTIE{texte=2004-06-01.0}-----
not switched off **until the end** of the war. the
so successful that **by the end** of the war, 63
-----PARTIE{texte=2004-06-20.0}-----
pale where he had his **headquarters** during the war.
-----PARTIE{texte=2004-06-29.0}-----
and germans but **as a supporter** of the war in
-----PARTIE{texte=2004-06-30.0}-----

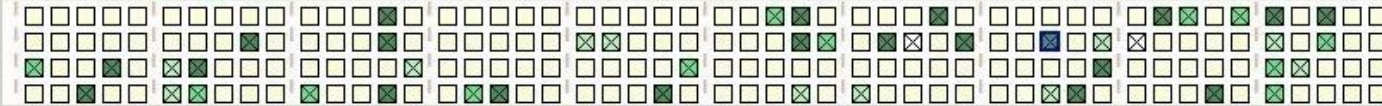
BI-TEXT TOPOGRAPHY (BBC)

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

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

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 :



Nb L. Sections sélectionnées : 0 N° Sect. : 1988:(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.
36

Position:<1832511>
Forme:<war>|Freq:1462
Lemme:<war>|Freq:1544
Cat:<NN_war>|Freq:1462

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

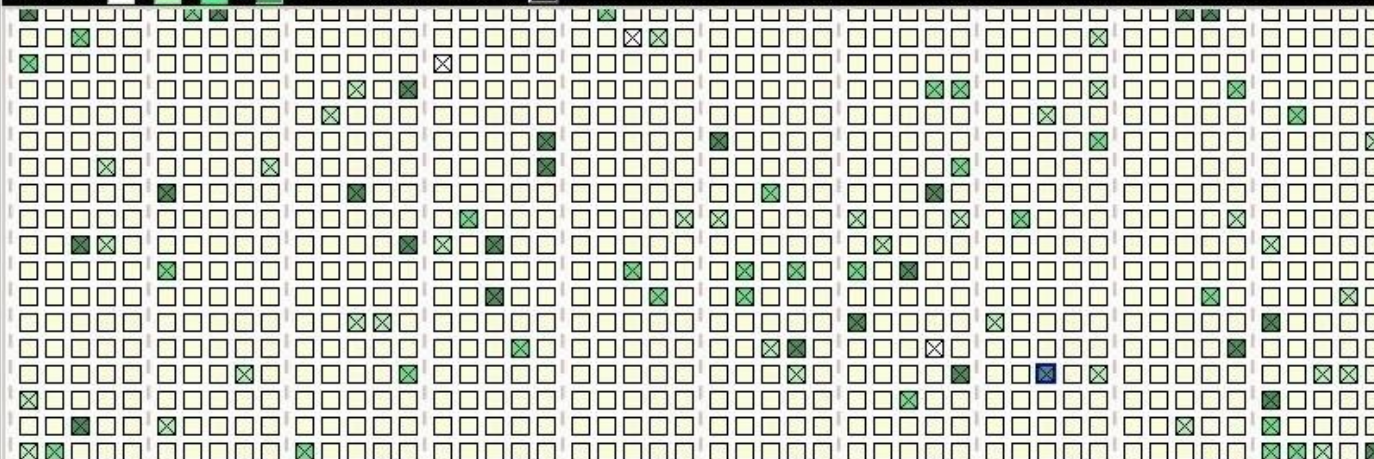
BI-TEXT TOPOGRAPHY (LENTA_RU)

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

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

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 :



Control-clic sur marqueur de page : sélection 5 sections | Shift-control-clic sur marqueur de page : sélection 25 sections (1 ligne)

Nb L. Sections sélectionnées : 0 N° Sect. : 1988:(719326,719716) Annotation : 1 Aperçu : 50

1 -->vip.lenta.ru -->что буш грядущий нам готовит?04.11.2004перспективы российско-американских отношений в ближайшие четыре года
2 президент ша 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 bbs 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 пообещал
7 осуществить налоговую реформу и сократить поток судебных тяжб 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 пообещал, что он будет налаживать отношения с
партнерами по нато и евросоюзу, с которыми у ша существовали противоречия по поводу войны в ираке.

Position:<719706>
Forme:<войны>|Freq:224
Lemme:<война>|Freq:363
Cat:<S_война>|Freq:325

CONCLUSIONS

- Combing systemic functional approach and textometric analysis offers new perspectives for bi-text alignment.
- Following a 'lexicogrammar' approach, textometric analysis reveals **significantly overrepresented lexicogrammatical patterns** in corresponding text zones (corpus parts).
- Extended lexicogrammatical patterns can be used to identify constituent parts of ST and its network of relations with TT (equivalence of function in context).

REFERENCES (1/2)

[CHO, 2009], *LEXICOMETRICA 3 “CORPUS MULTILINGUES”*. ONLINE PUBLICATION <<http://Lexicometrica.Univ-paris3.Fr/Numspeciaux/Special8.Htm>>

[Fleury, 2013], *LE TRAMEUR. PROPOSITIONS DE DESCRIPTION ET D'IMPLÉMENTATION DES OBJETS TEXTOMÉTRIQUES*. ONLINE PUBLICATION: <<http://www.tal.univparis3.fr/trameur/trameur-propositions-definITIONS-objets-textometriques.pdf>>

[Halliday, 1992], *LANGUAGE THEORY AND TRANSLATION PRACTICE*. *RIVISTA INTERNAZIONALE DI TECNICA DELLA TRADUZIONE* 0, 15-25.

[Harris, 1988], *BI-TEXT, A NEW CONCEPT IN TRANSLATION THEORY*. *LANGUAGE MONTHLY*, 54: PP 8-10.

[Lamalle C., Fleury S. & Salem A., 2006], *VERS UNE DESCRIPTION FORMELLE DES TRAITEMENTS TEXTOMÉTRIQUES*. *ACTES DES 8ÈMES JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES (JADT'06)*, pp. 583-593

[Manfredi, 2008], *TRANSLATING TEXT AND CONTEXT: TRANSLATION STUDIES AND SYSTEMIC FUNCTIONAL LINGUISTICS. VOL. 1: TRANSLATION THEORY*. *QUADERNI DEL CESLIC: FUNCTIONAL GRAMMAR STUDIES FOR NON-NATIVE SPEAKERS OF ENGLISH*. BOLOGNA, DUPRESS.

REFERENCES (2/2)

[Manfredi, 2011], *SYSTEMIC FUNCTIONAL LINGUISTICS AS A TOOL FOR TRANSLATION TEACHING: TOWARDS A MEANINGFUL PRACTICE. RIVISTA INTERNAZIONALE DI TECNICA DELLA TRADUZIONE* 13, 49-62.

[Salem, 1987], *PRATIQUE DES SEGMENTS RÉPÉTÉS. ESSAI DE STATISTIQUE TEXTUELLE. PARIS, KLINCKSIECK.*

[Salem, 2004], *INTRODUCTION À LA RÉSONANCE TEXTUELLE. ACTES DES 7ES JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES (JADT'04), PP. 986-992.*

[Steiner & Yallop, 2001] (EDS) : *EXPLORING TRANSLATION AND MULTILINGUAL TEXT PRODUCTION : BEYOND CONTENT, BERLIN/NEW YORK, MOUTON DE GRUYTER.*

[Schmid, 1994], *PROBABILISTIC PART-OF-SPEECH TAGGING USING DECISION TREES. PROCEEDINGS OF INTERNATIONAL CONFERENCE ON NEW METHODS IN LANGUAGE PROCESSING, PP. 44-49.*

[Klementiev & Roth, 2006], *WEAKLY SUPERVISED NAMED ENTITY TRANSLITERATION AND DISCOVERY FROM MULTILINGUAL COMPARABLE CORPORA. ACL-44, PP. 817-824.*

[Zimina, 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.*