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 translatologists 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. […]

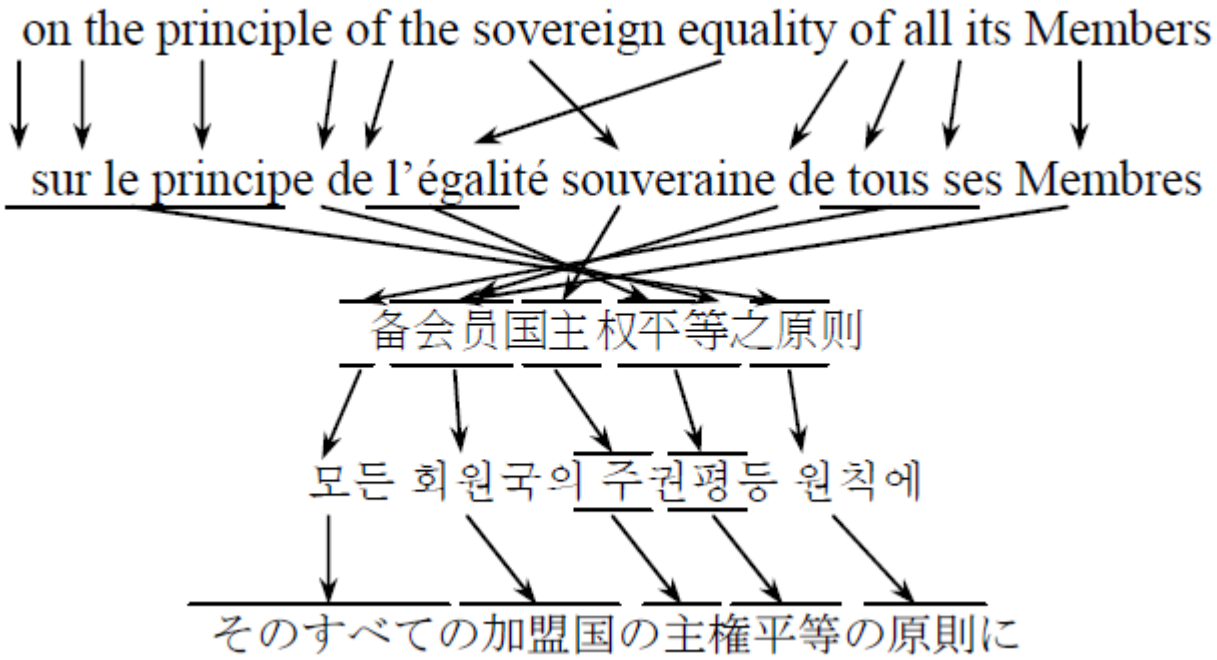
To semioticians, by the way, I submit that *bi-text* falls into the same paradigm as *intertext,* in that it is a construct of two or more related texts."

Bi-text, a new concept in translation theory.

Brian Harris, 1988

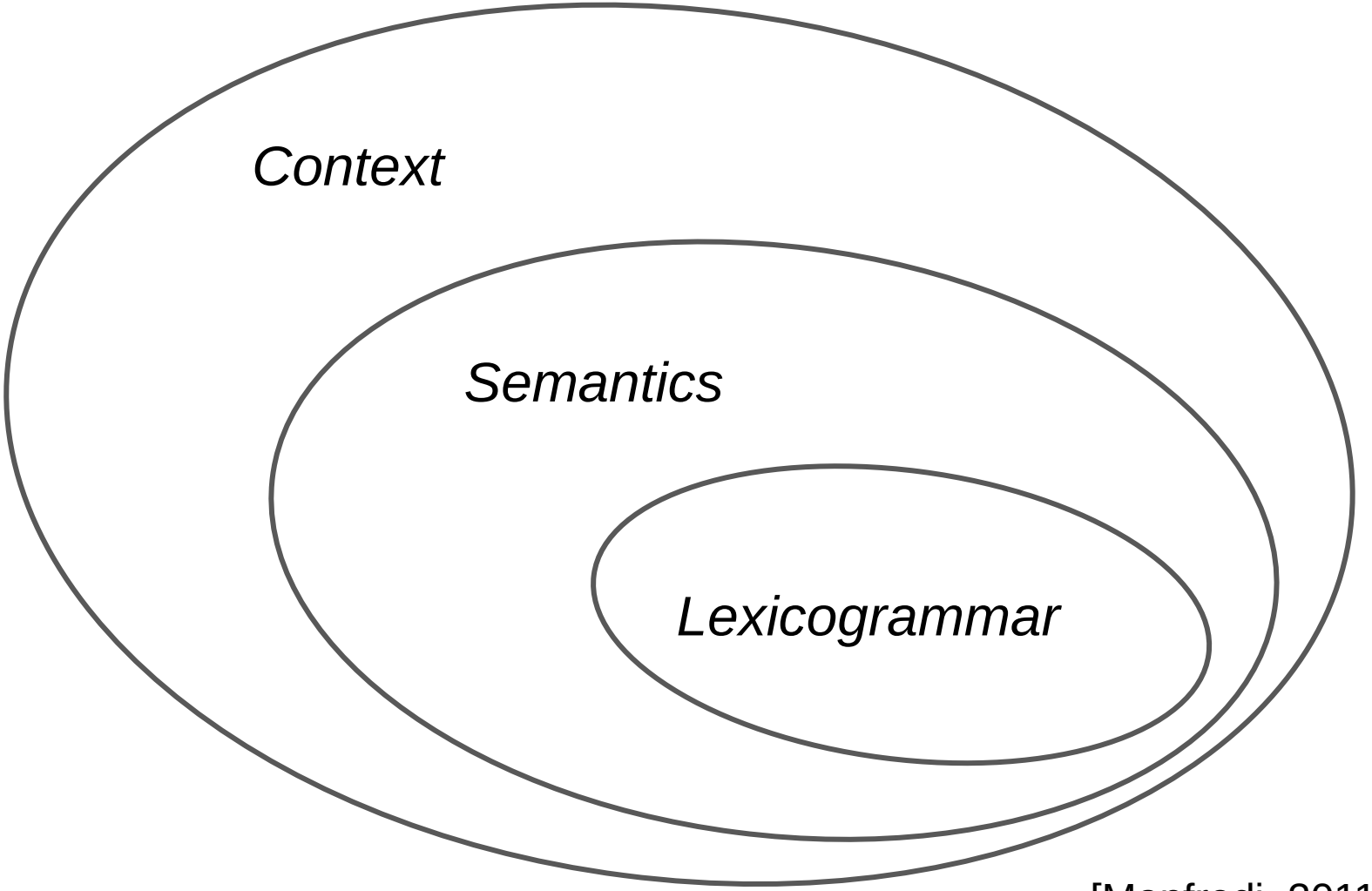
HOW TO ALIGN MULTILINGUAL TEXTS?

"configurations of multidimensional meanings, rather than […] containers of *content*" [Steiner and Yallop, 2001]



[CHO, 2009]
Lexicometrica, vol. 3 "Corpus multilingues"

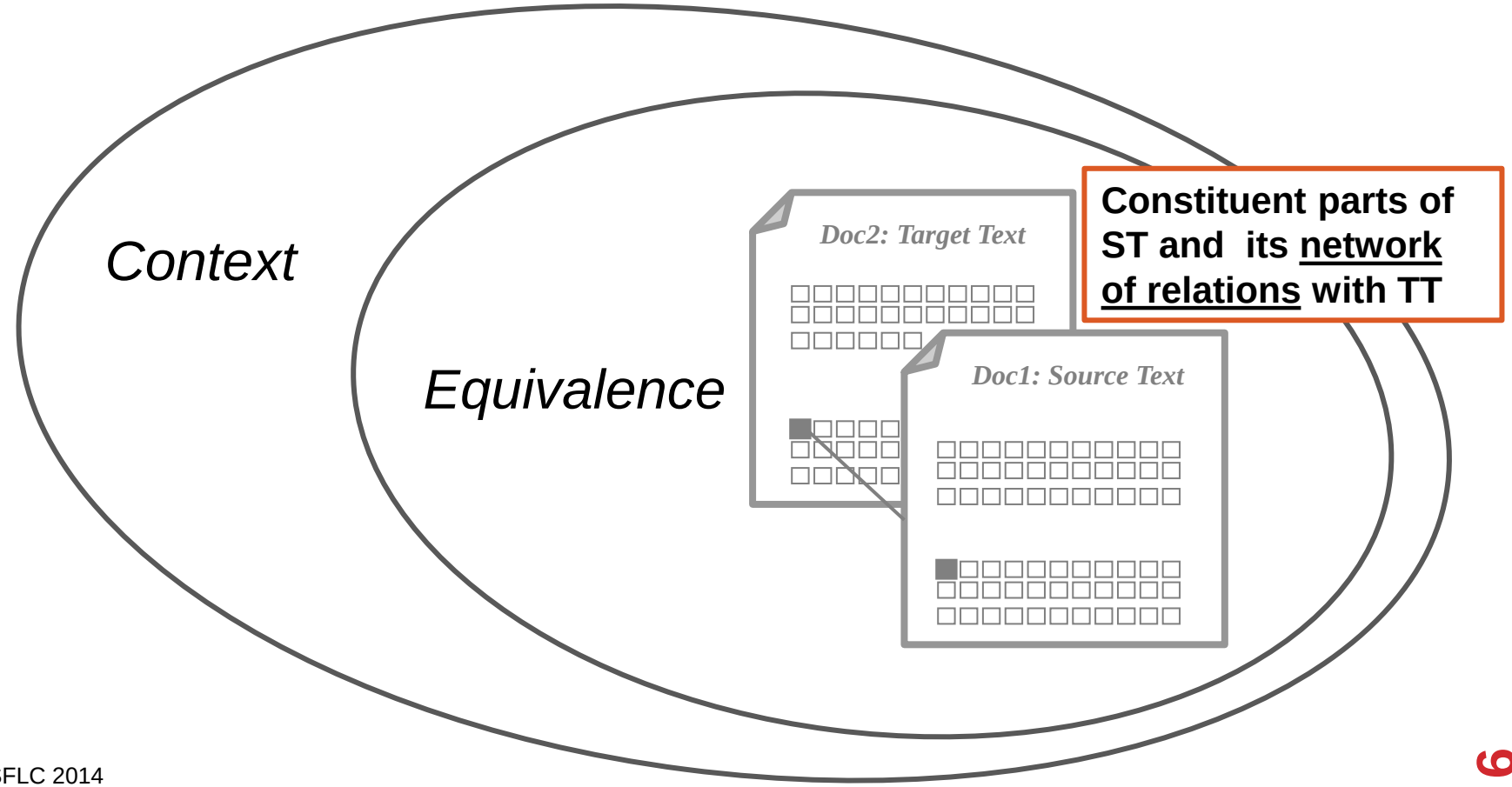
SFL: ANALYTICAL TOOL FOR BI-TEXT ANALYSIS



[Manfredi, 2011]

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
<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. /…/	<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. /…/

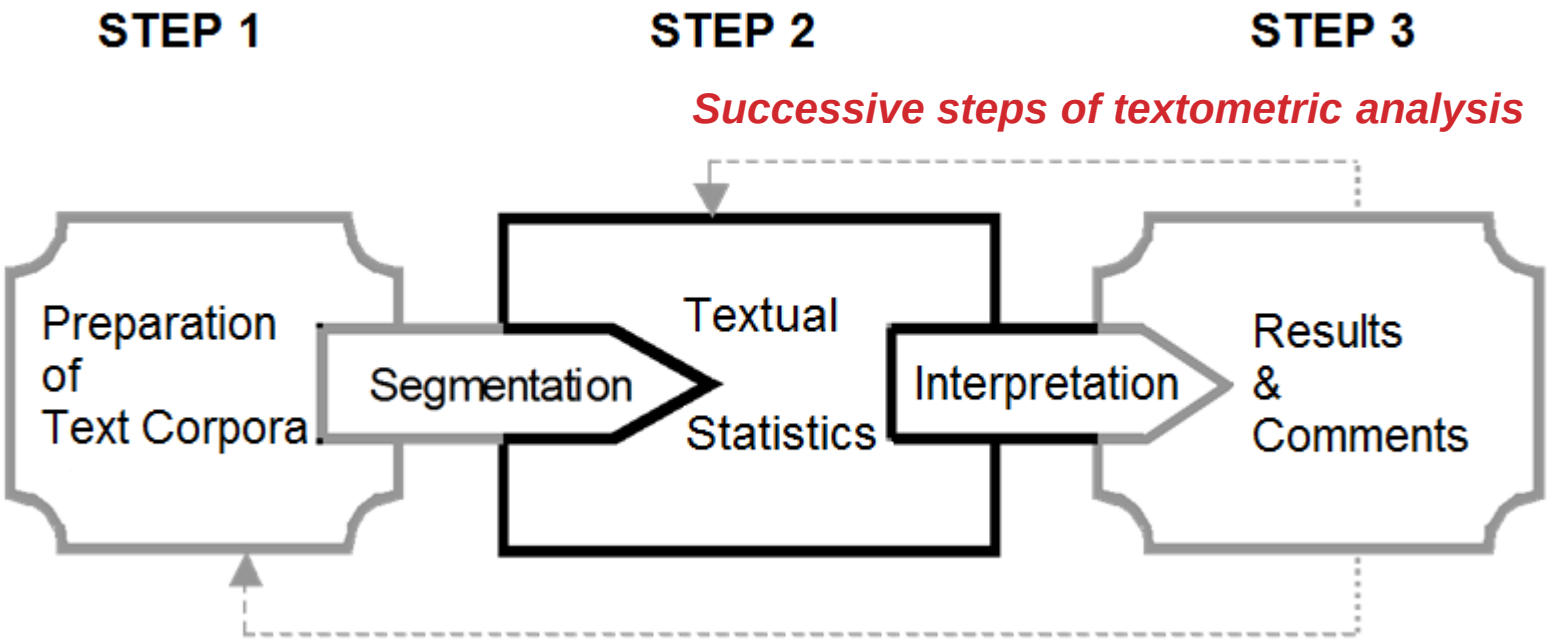
TEXTOMETRIC ANALYSIS (1/2)

Object of study:

- Text (or texts)

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

TEXTOMETRIC ANALYSIS (2/2)

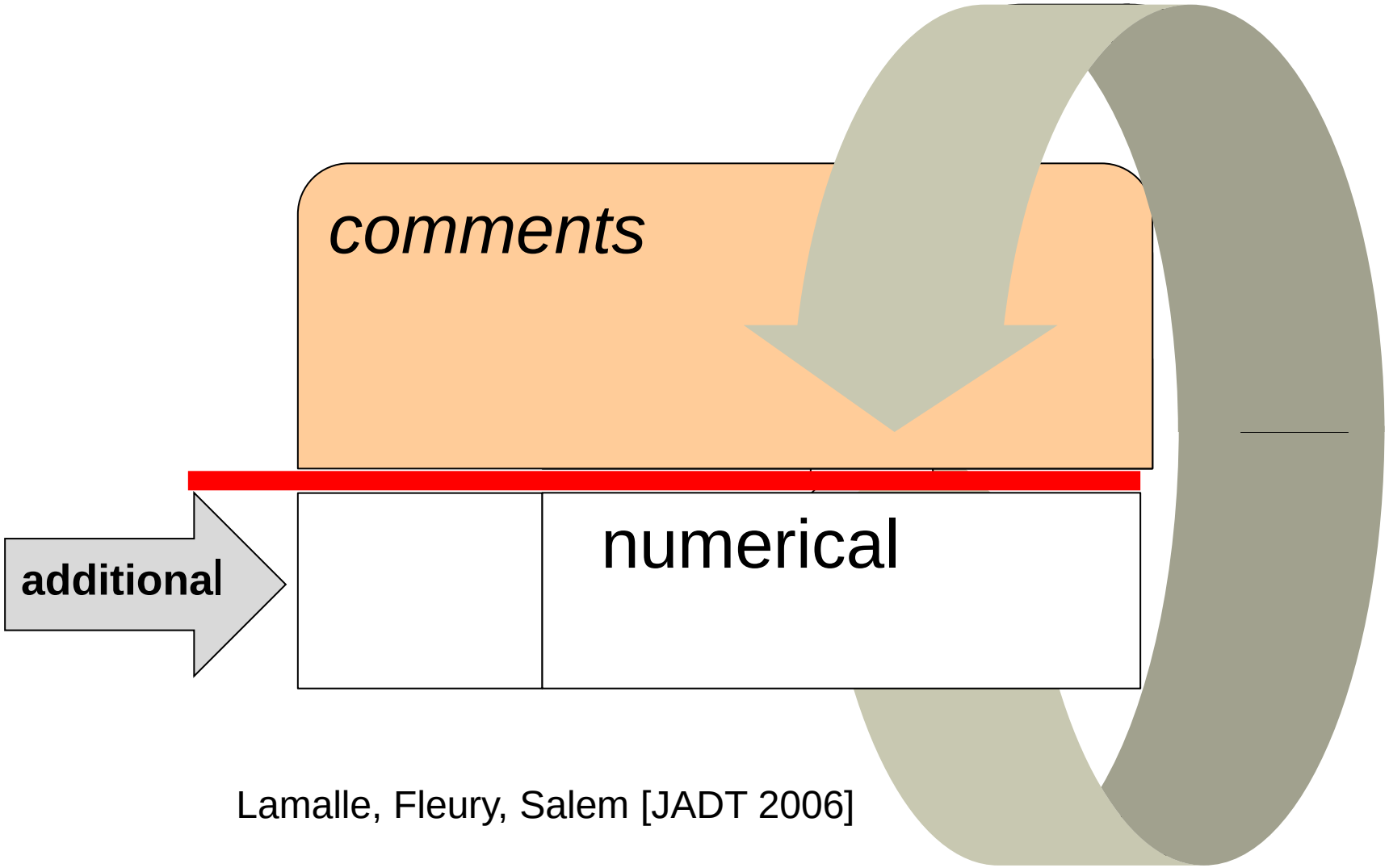


Lamalle, Fleury, Salem [JADT 2006]

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]

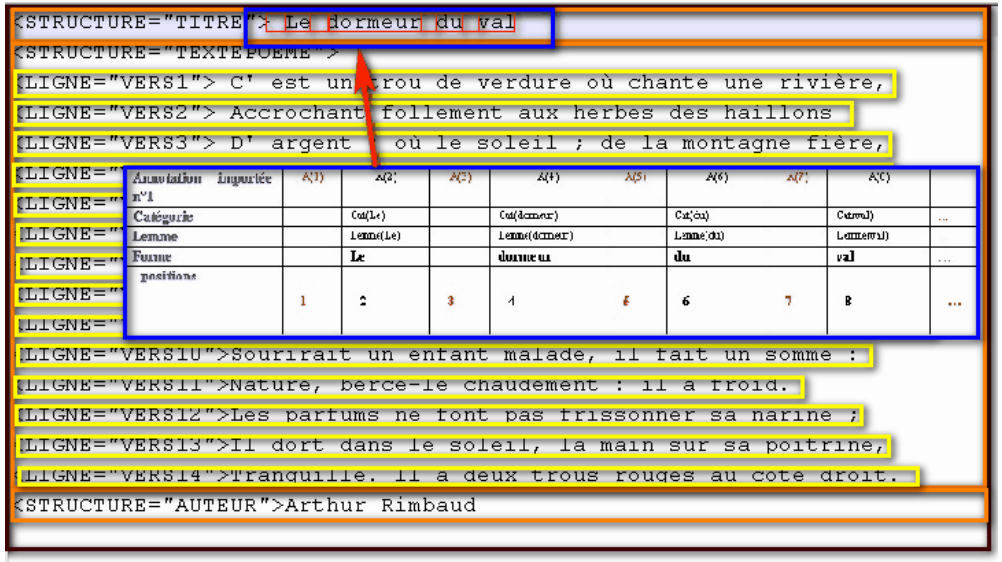
TEXT FLOW

Trameur [Fleury, 2013]

***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*.**



TEXT ANNOTATIONS

Trameur [Fleury, 2013]

Lemmatisation, morpho-syntactic tagging

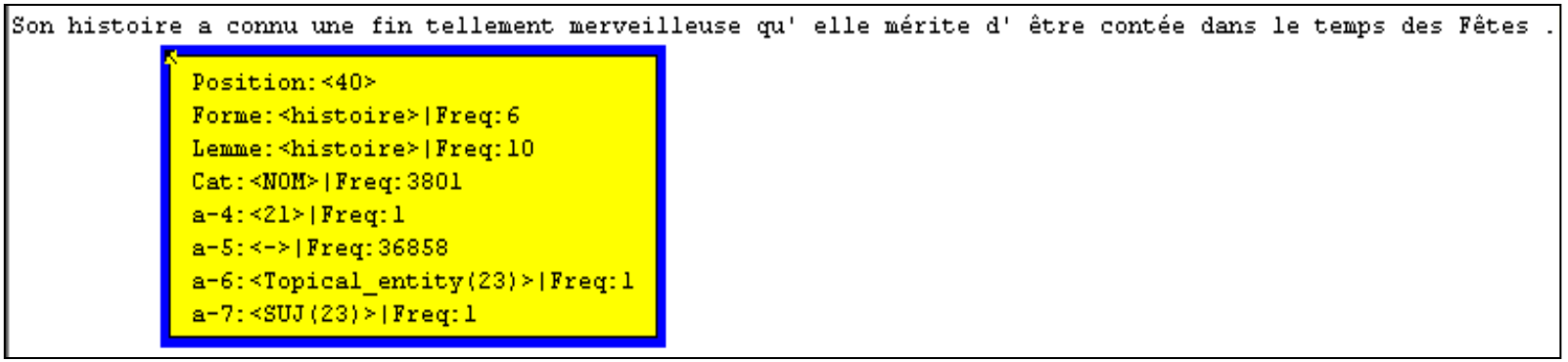
- Integrated into *Trameur* via *TreeTagger* [Schmid, 1994]

Semantic annotation, etc.

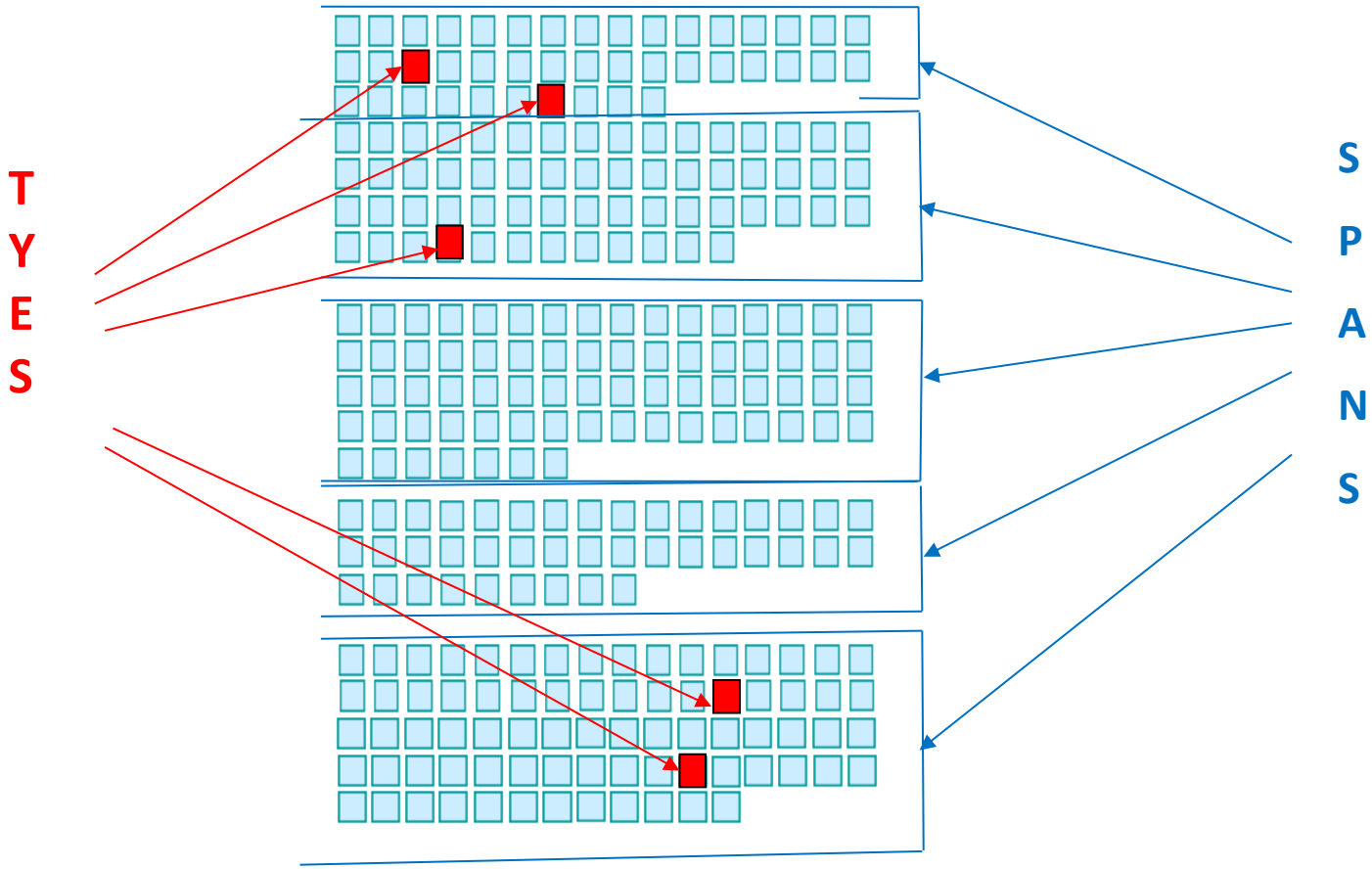
- Possibility to implement multi-level annotations on the *Thread*

Annotation of dependency relations

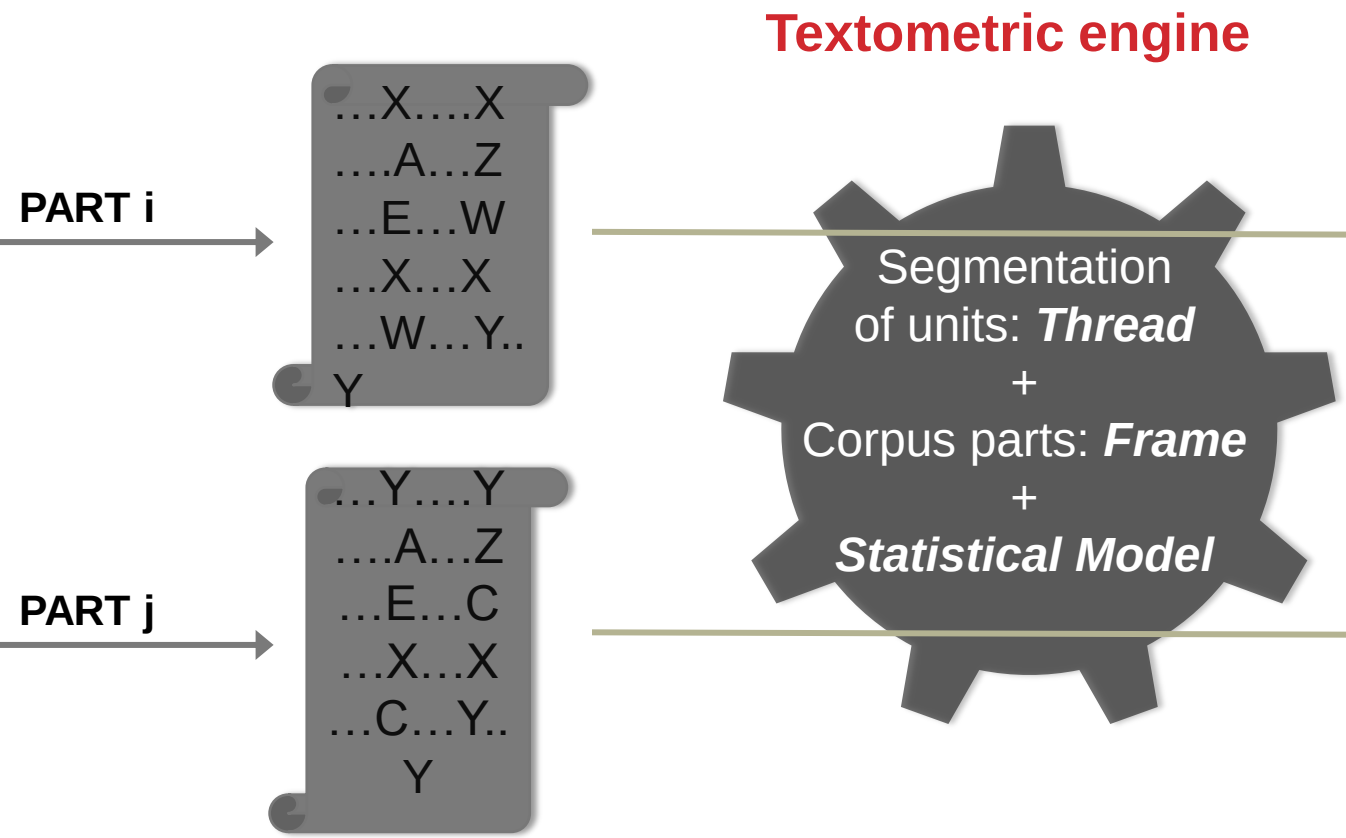
Thread = sequence of annotated items with position identifiers



STATISTICAL TABLES TYPES & SPANS



TRAMEUR [S. FLEURY, 2013]



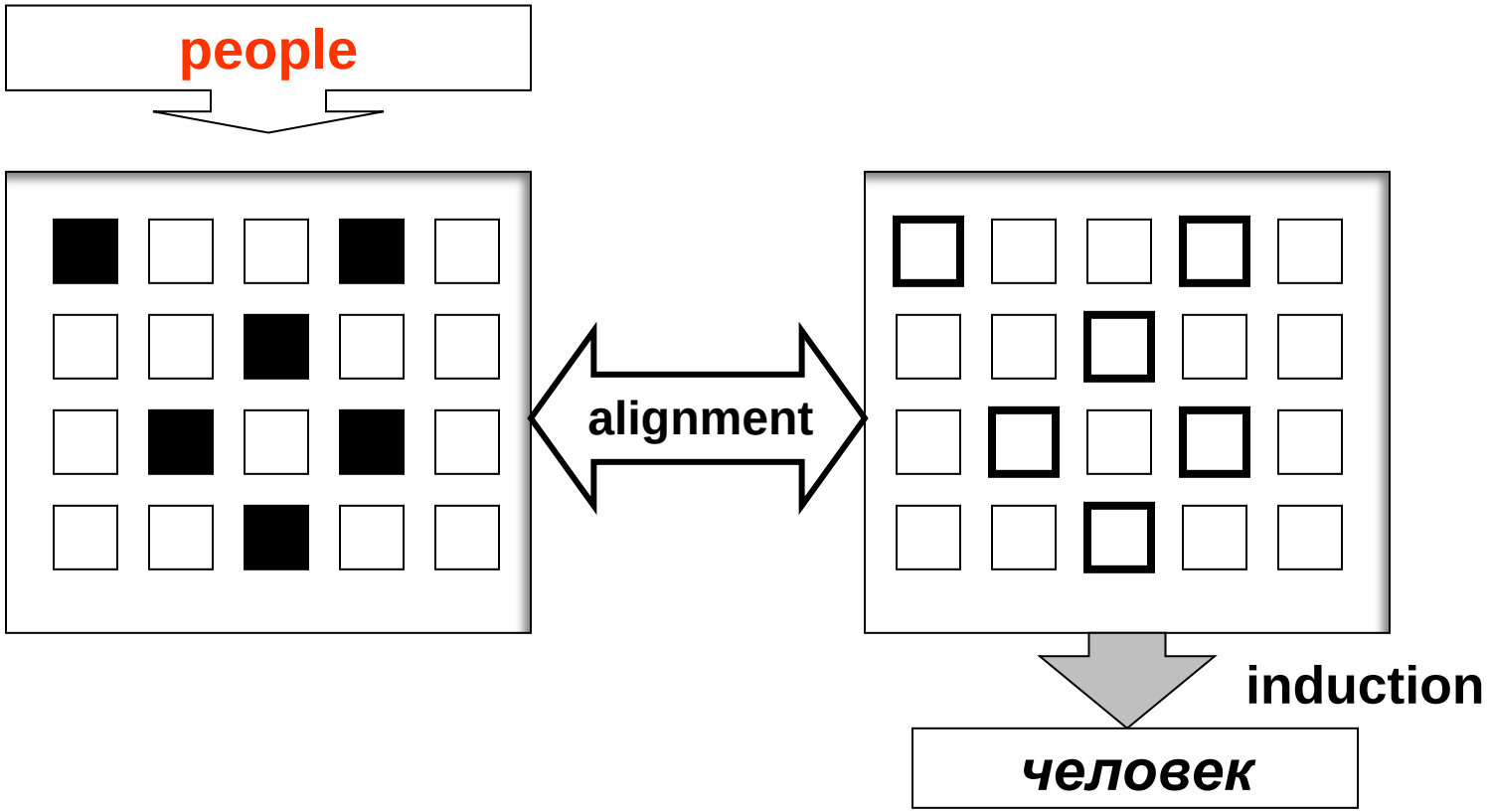
Token	Prob. Ind.
X	23.43
Y	12.68
Z	5.57
W	5.66

Token	Prob. Ind.
Y	13.73
X	21.86
A	7.75
C	6.55

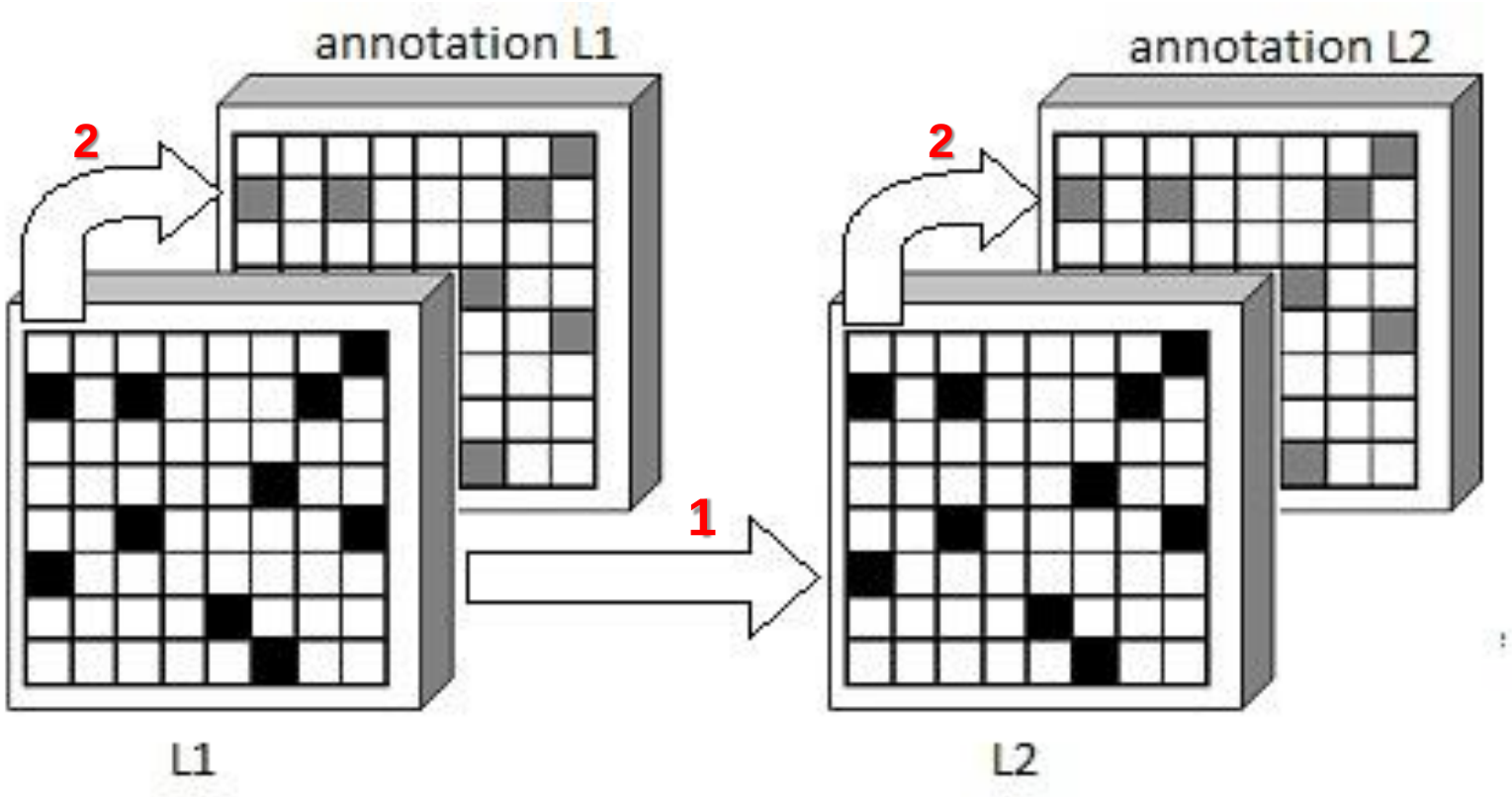
http://www.tal.univ-paris3.fr/trameur

TEXTUAL RESONANCE

Salem, Zimina [JADT 2004]



RESONANCE PROPAGATION TO ANNOTATIONS



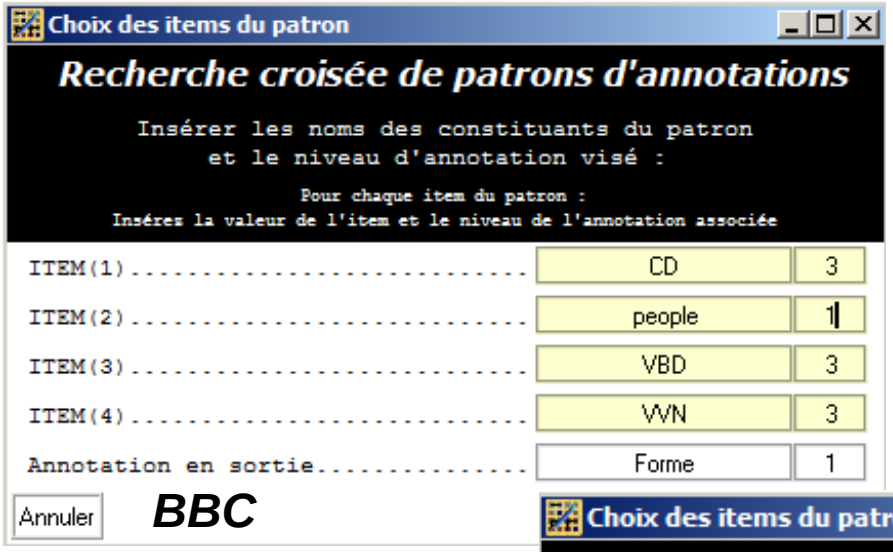
CROSS-ANNOTATIONS AND REPEATED SEGMENTS

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

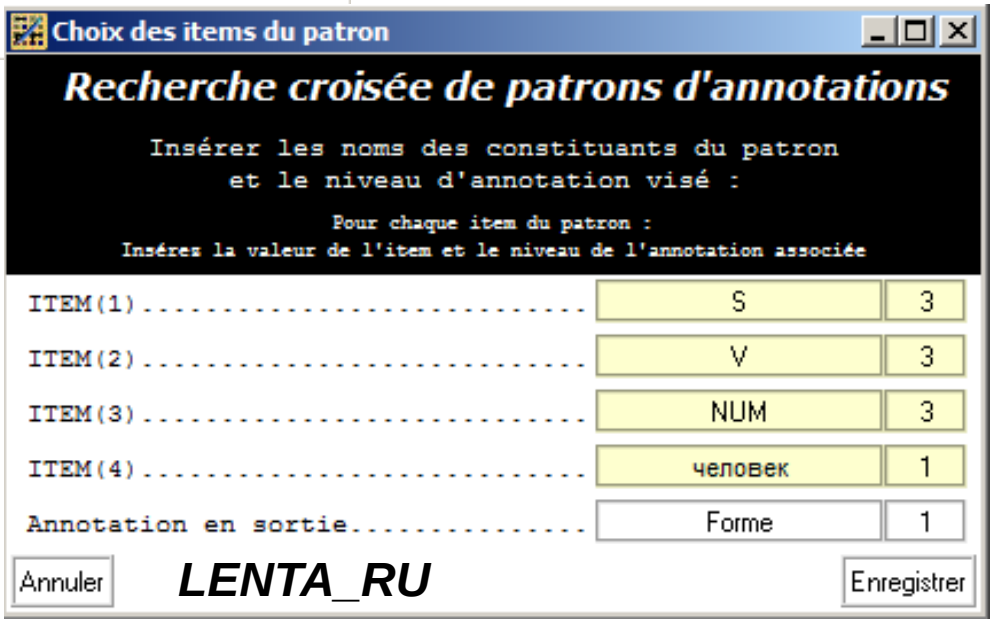


CD + **people** + **VBD** + **VVN**

(English)

S + **V** + **NUM** + **человек**

(Russian)



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**

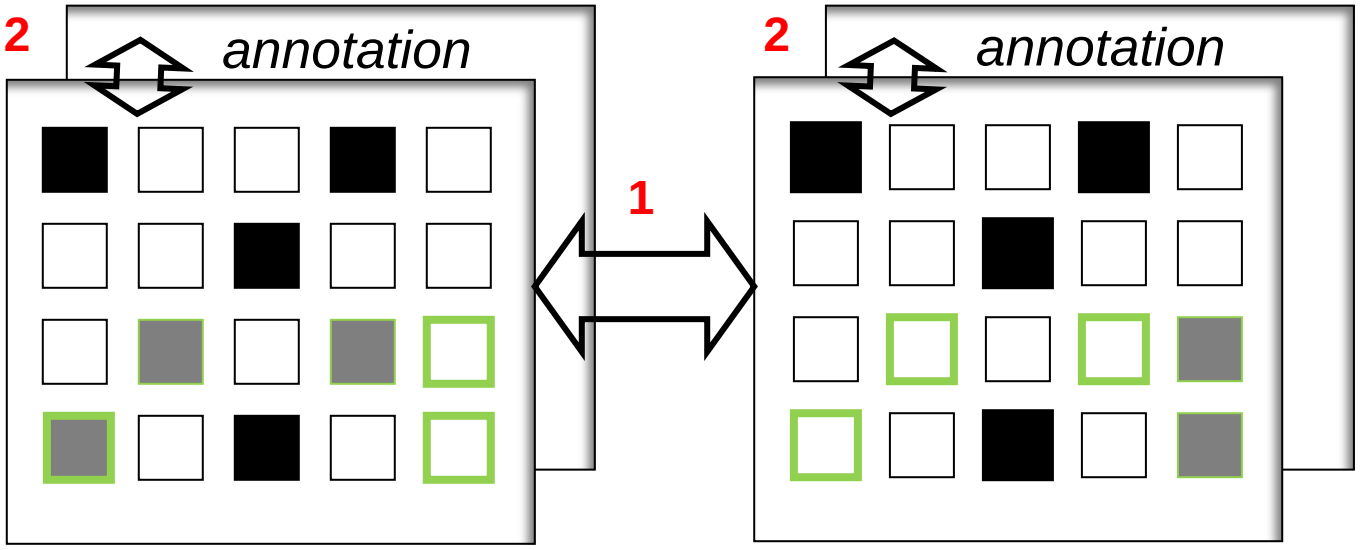
----------BBC=2002-06-17.1----------
general on friday in which **12 people were killed** and
----------LENTA_RU=2002-06-17.1----------
взрывчаткой машина. в результате **теракта погибли 12 человек**

ITERATIVE INDUCTION

Contextually conditioned variants

researchers ⟹ *учёны(е|х)*

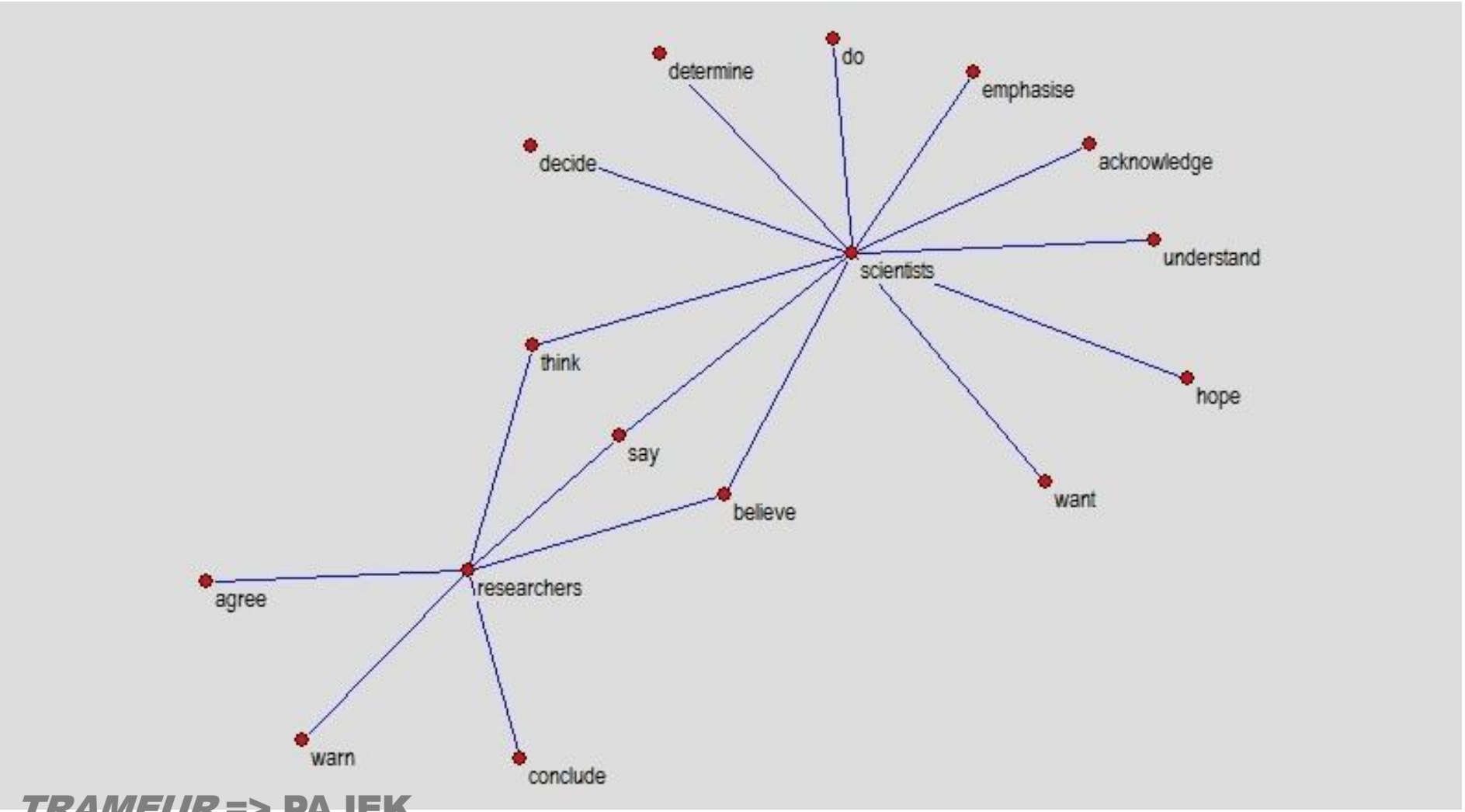
scientists ⟸⟹ *исследовател(и|ей)*



[Zimina, 2004]

BBC–LENTA_RU [EN]

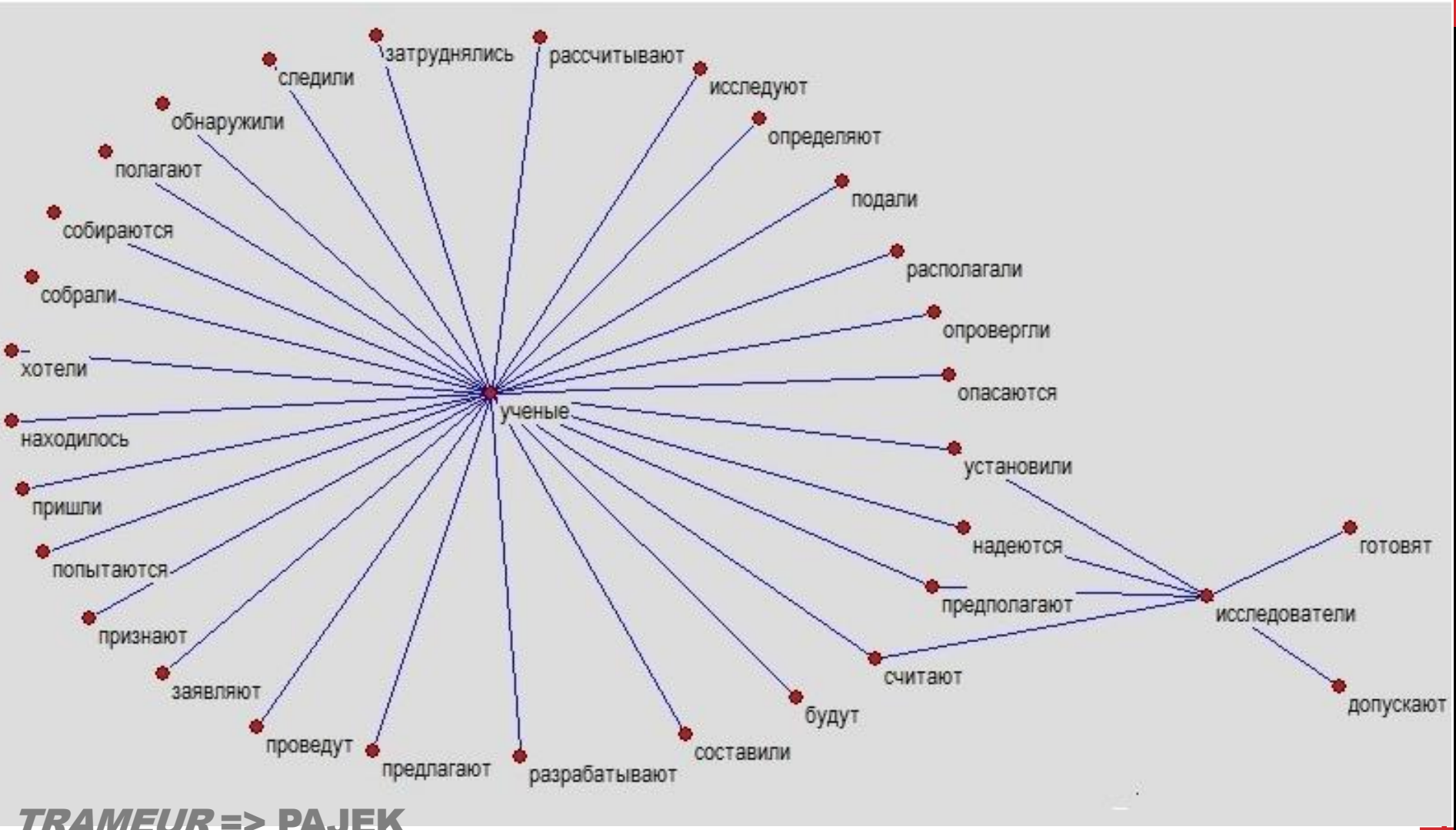
RESEARCHERS / SCIENTISTS + VVP [F = 72]



***TRAMEUR* => PAJEK**

BBC–LENTA_RU [RU]

УЧЁНЫЕ / ИССЛЕДОВАТЕЛИ + V [F = 98]



***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

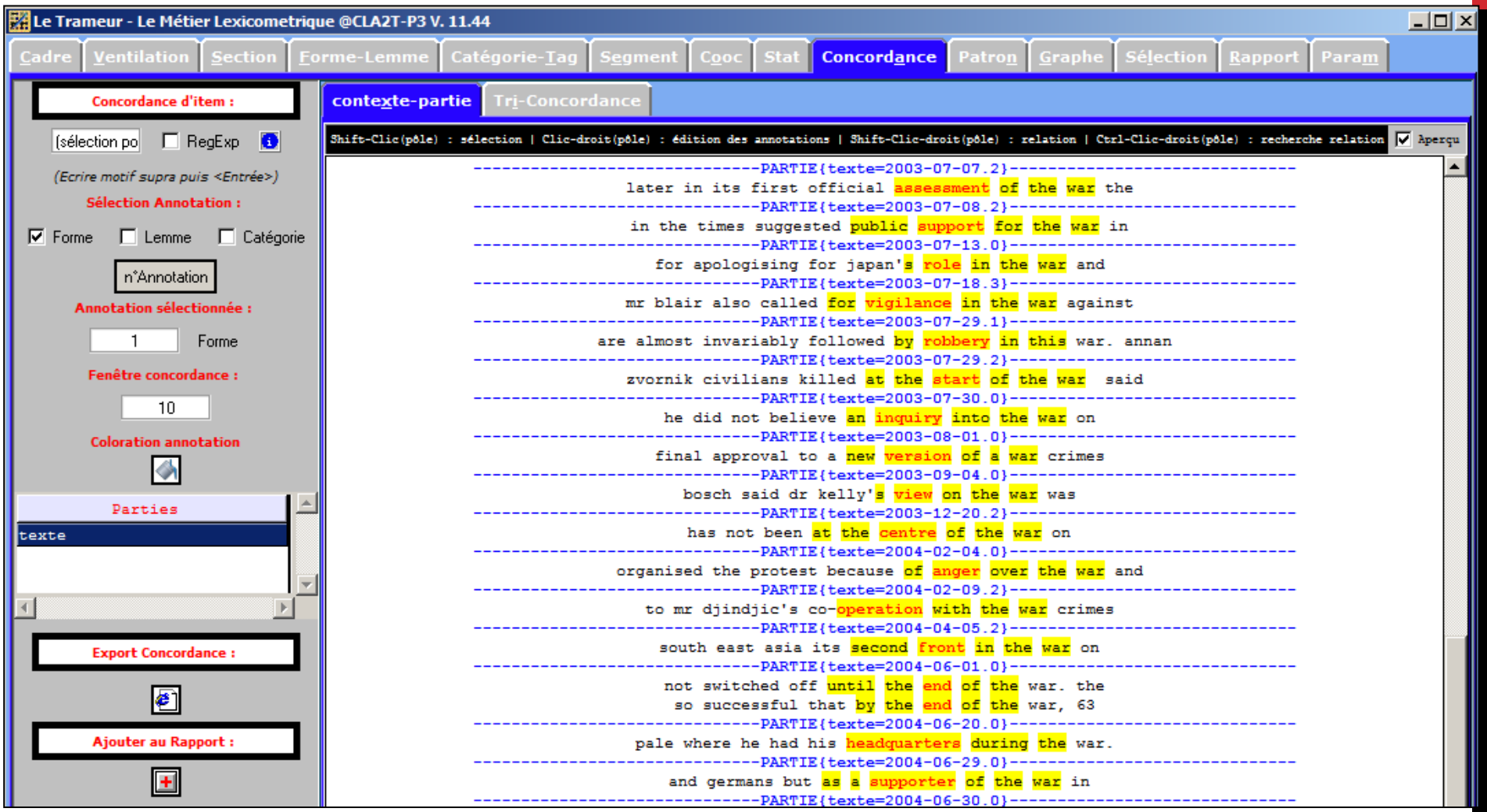
----------PARTIE{texte=2002-09-07.2}----------
light-years away. **scientists hope** this will enable them to test one of einstein's assumptions
в субботу **ученые проведут** эксперимент, призванный доказать или опровергнуть общую теорию

----------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,

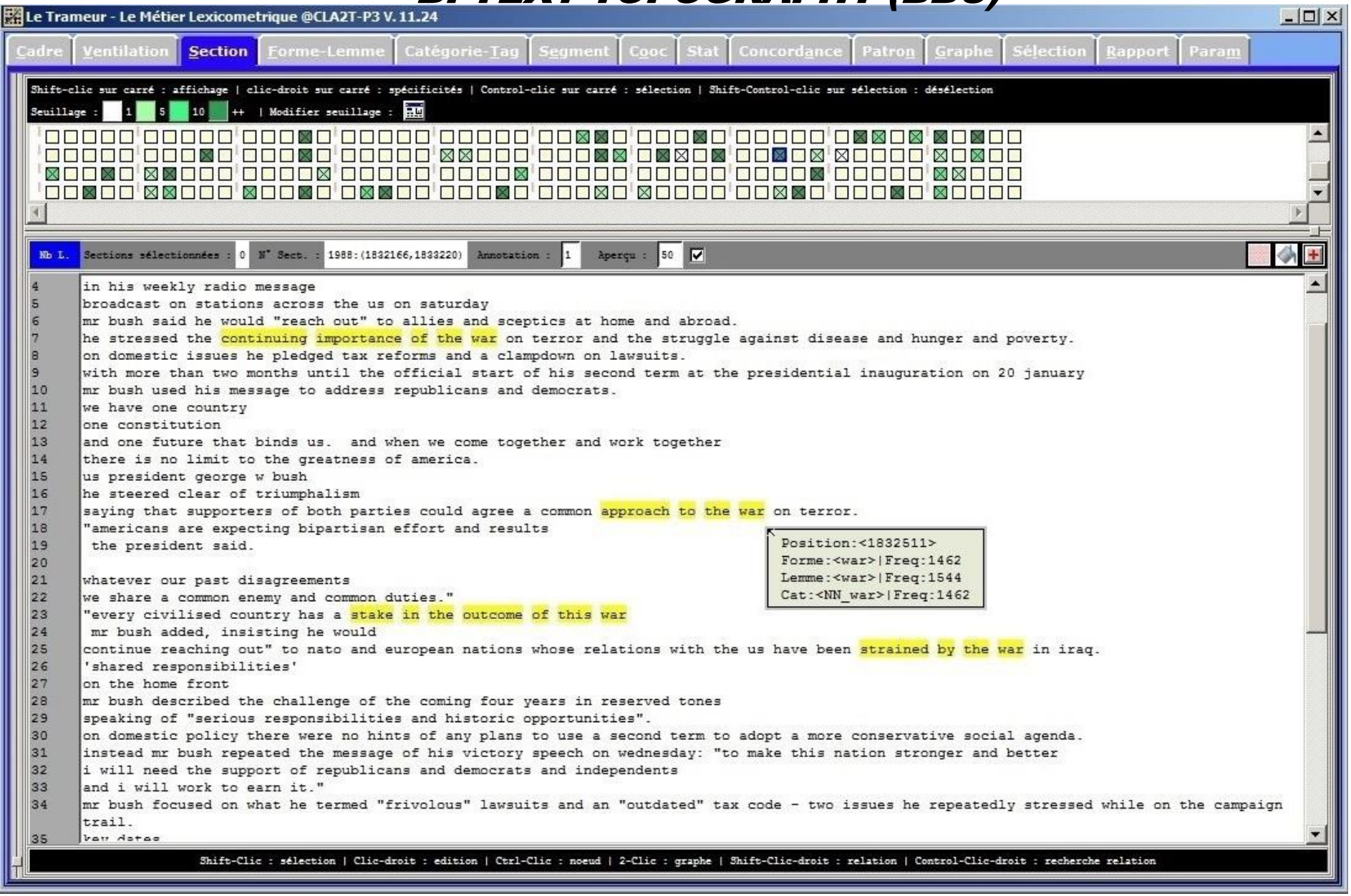
----------PARTIE{texte=2004-02-27.2}----------
some **researchers think** comet impacts may even have seeded the early earth with the chemistry
некоторые **исследователи надеются** найти на ядре подтверждение теории о космическом

TEXTOMETRIC INVENTORIES AND LEXICOGRAMMAR

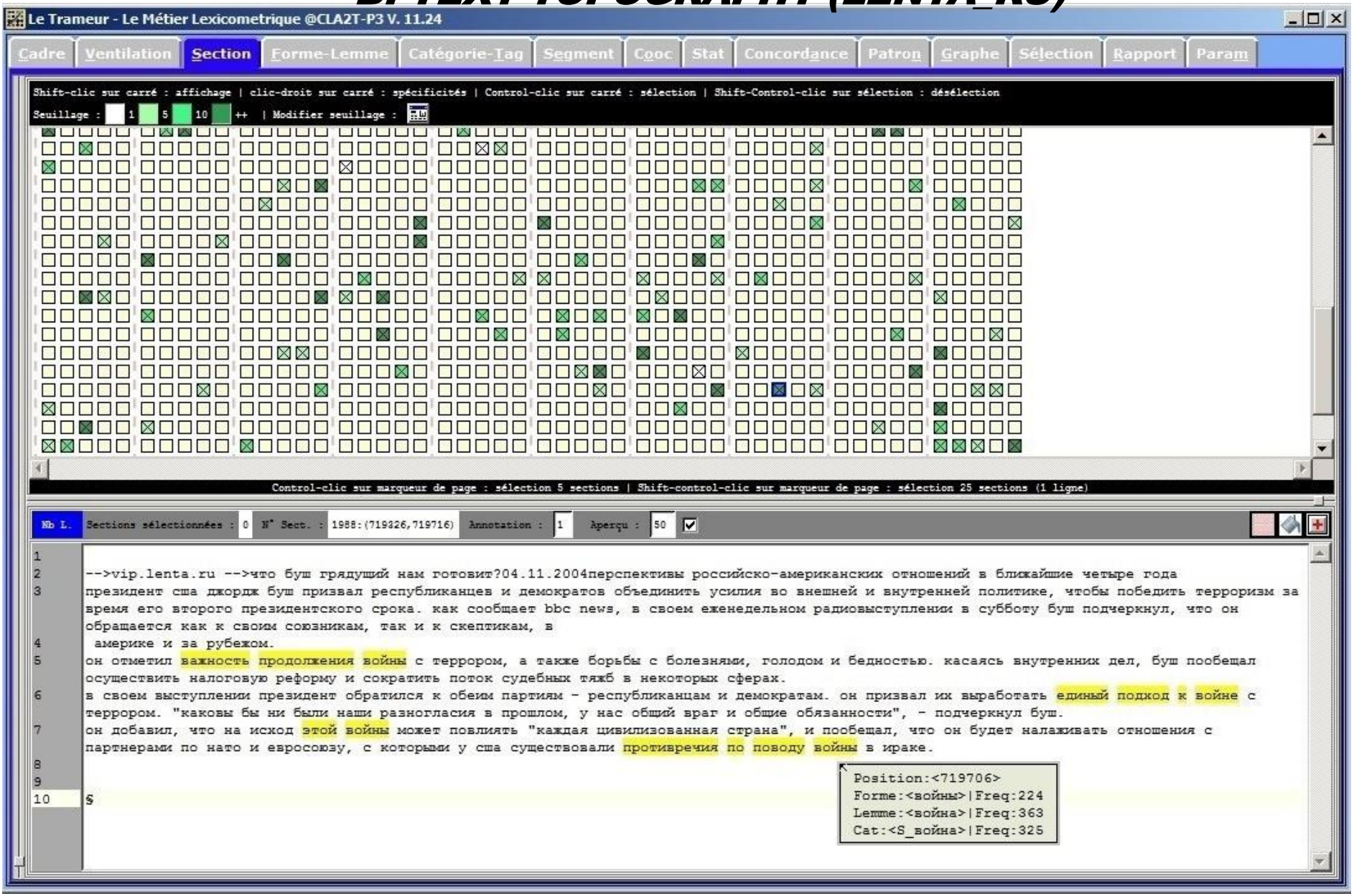
NOUN + *PREPOSITION* + *DETERMINER* + *WAR* (English)



BI-TEXT TOPOGRAPHY (BBC)



BI-TEXT TOPOGRAPHY (LENTA_RU)



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.*