Syntactic Parsing versus MWEs: What can fMRI signal tell us

Murielle Fabre, Yoann Dupont, Éric Villemonte de la Clergerie

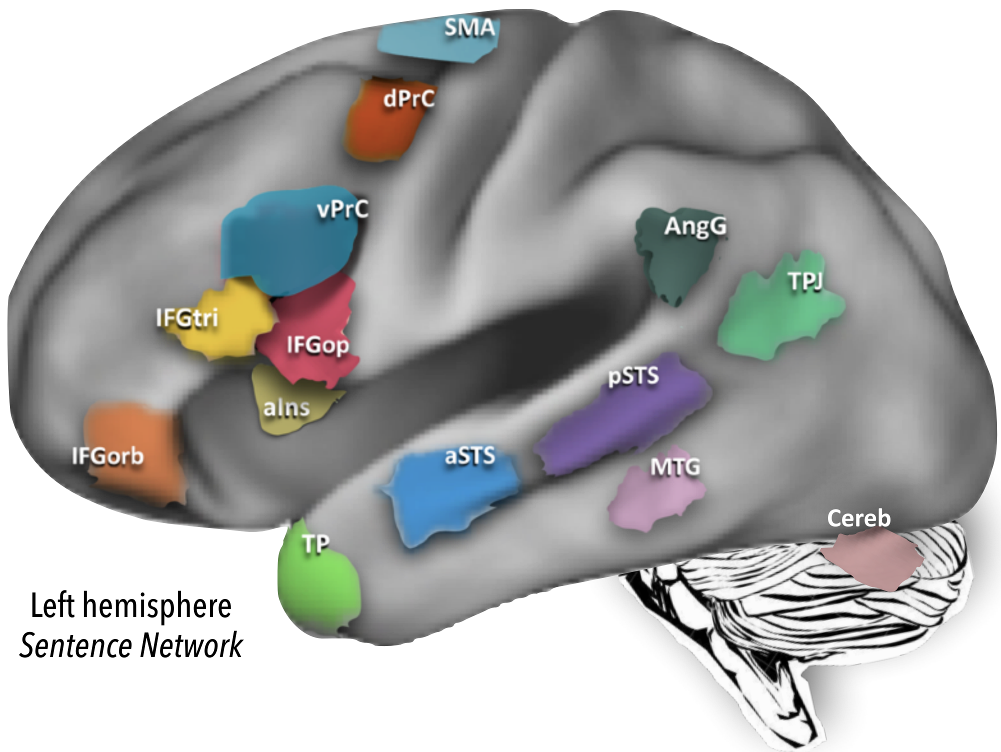
Syntactic Parsing versus MWEs

what can fMRI signal tell us

Murielle Fabre, Yoann Dupont, Eric de la Clergerie

Project & Approach

Bring together
computational linguistics and cognitive neuro-imaging
to shed light on sentence comprehension and its neural bases

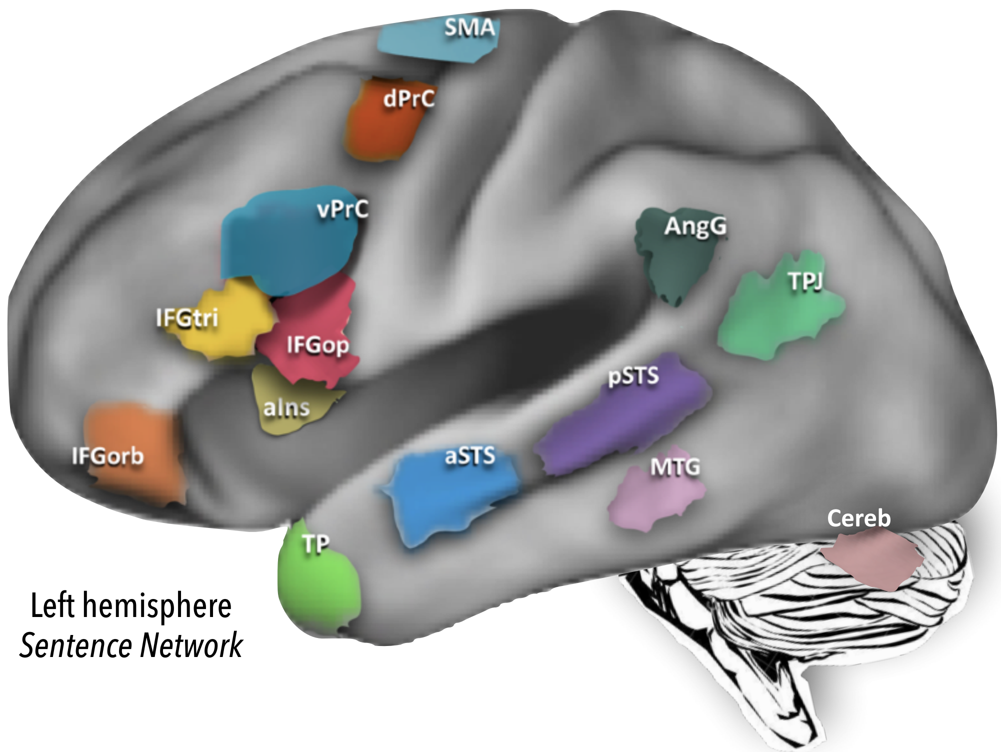


Left hemisphere
Sentence Network

Project & Approach

Bring together
computational linguistics and cognitive neuro-imaging
to shed light on sentence comprehension and its neural bases

Does MWE processing pattern together with sentence-structure building effects in the brain?



Naturalistic Corpus : The Little Prince in 3 languages

Il y a six ans déjà que mon ami s'en est allé avec son mouton. Si j'essaie ici de le décrire, c'est afin de ne pas l'oublier. C'est triste d'oublier un ami.

French

Six years ago that my friend left wit his sheep. If I try to describe him, it's ignorer not to forget him. It is sad to forget a friend.

English

我的朋友找他的羊已离开六年了。
我在描述他,是了不忘他。
把朋友忘了是一件酸的事。

Chinese

Naturalistic Corpus : The Little Prince in 3 languages

Il y a six ans déjà que mon ami s'en est allé avec son mouton. Si j'essaie ici de le décrire, c'est afin de ne pas l'oublier. C'est triste d'oublier un ami.

French

Six years ago that my friend left wit his sheep. If I try to describe him, it's ignorer not to forget him. It is sad to forget a friend.

English

我的朋友找他的羊已离开六年了。
我在描述他，是了不忘他。
把朋友忘了是一件酸的事。

Chinese

"*Everyday listening*" conditions in English

Participants (51)
American native speakers
(32 women, 18-37 years old)

Task: Listen to the audiobook
The Little Prince (1 h 38 min) 9 runs
+ Comprehension questions after each run

Recoding : 3T MRI scanner 32-channel head coil at the Cornell MRI Facility. Muti-echo sequence.
Preprocessing : FSL, AFNI + the signal-to-noise ratio, using multi-echo independent components analysis (ME-ICA) (Kundu et al., 2013)

Analysis : General Linear Model (GLM - SPM12)
Control Regressors : pitch (f0), acoustic volume (RMS), Word rate, Word frequency.

The Team for this study

Murielle Fabre

Yoann Dupont

Eric de la Clergerie

Cornell University

UNIVERSITE D'ORLEANS

USPC Université Sorbonne Paris Cité

inria inventeurs du monde numérique

Shohini Bhattasali

Wenming Luh

John Hale

Mathieu Constant

Hazem Al-saied

cnrs

ATILF UMR 7118
(CNRS/Université de Lorraine)

Christophe Pallier

Inserm

UNICOG Cognitive Neuroimaging Lab

Road map for today

A Parsing vs. MWEs

B Identifying MWEs

C Stability of PMIs

D fMRI results

E Next steps

Parsing versus MWEs

A

Scientific Questions and Hypotheses

Multi-word expressions

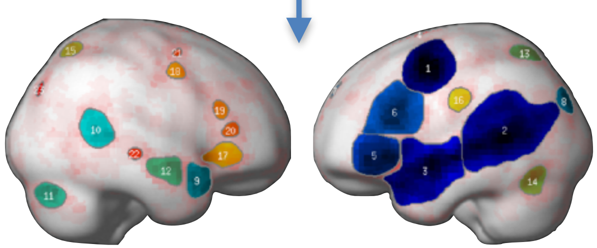
Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

Scientific Questions and Hypotheses

Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

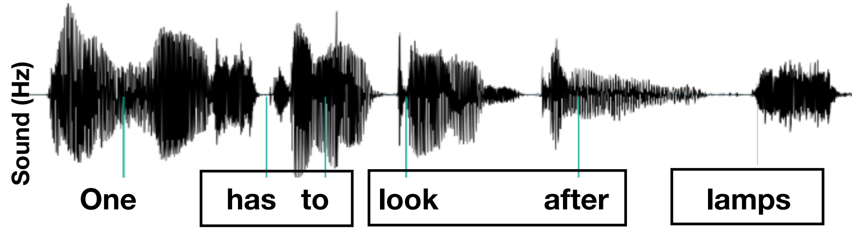
MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.



MWE Processing

Structure building

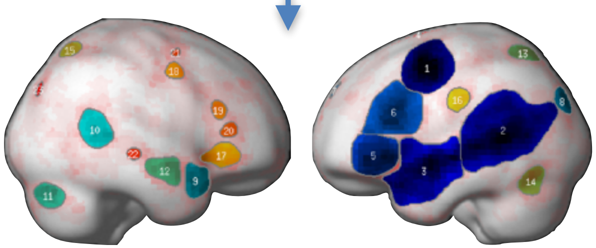
Scientific Questions and Hypotheses



Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

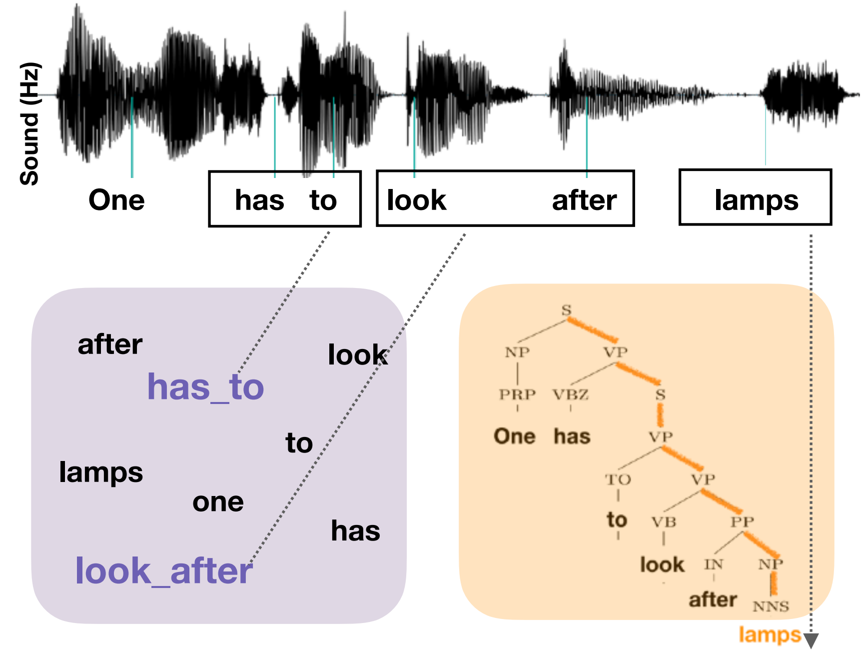
MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.



MWE Processing

Structure building

Scientific Questions and Hypotheses



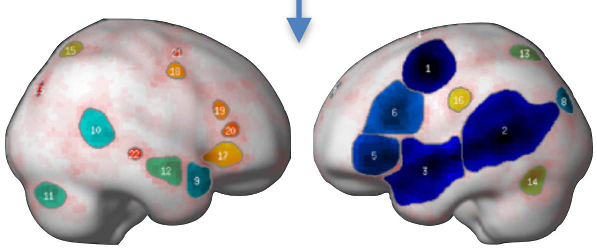
MWE Processing

Structure building

Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.

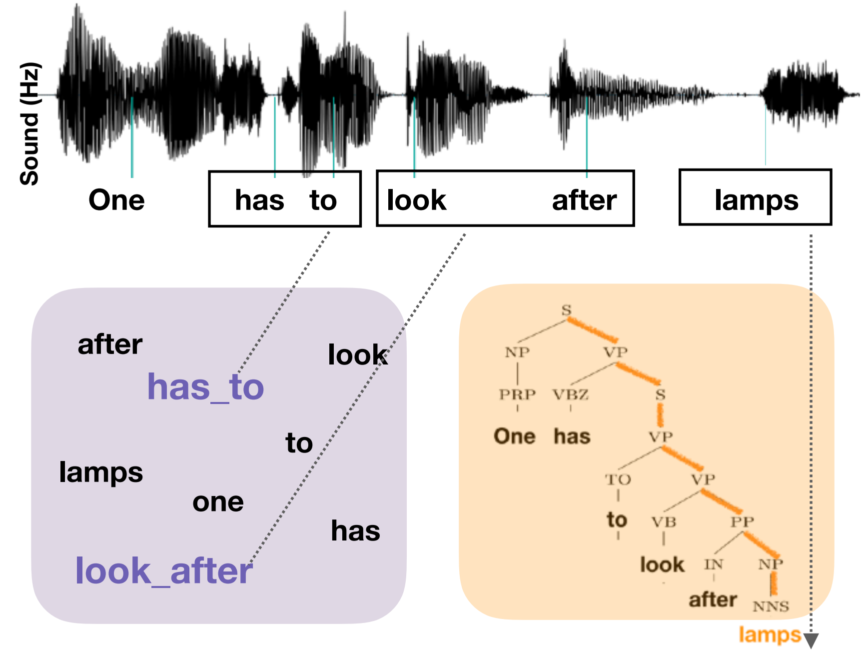


Scientific Questions and Hypotheses

Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.

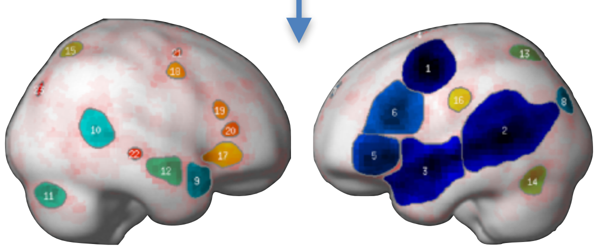


MWE Processing

a computational graded quantification identifying expressions likely to be processed as units, rather than built-up compositionally

Structure building

a computational measure tracking tree-building work needed in composed syntactic phrases

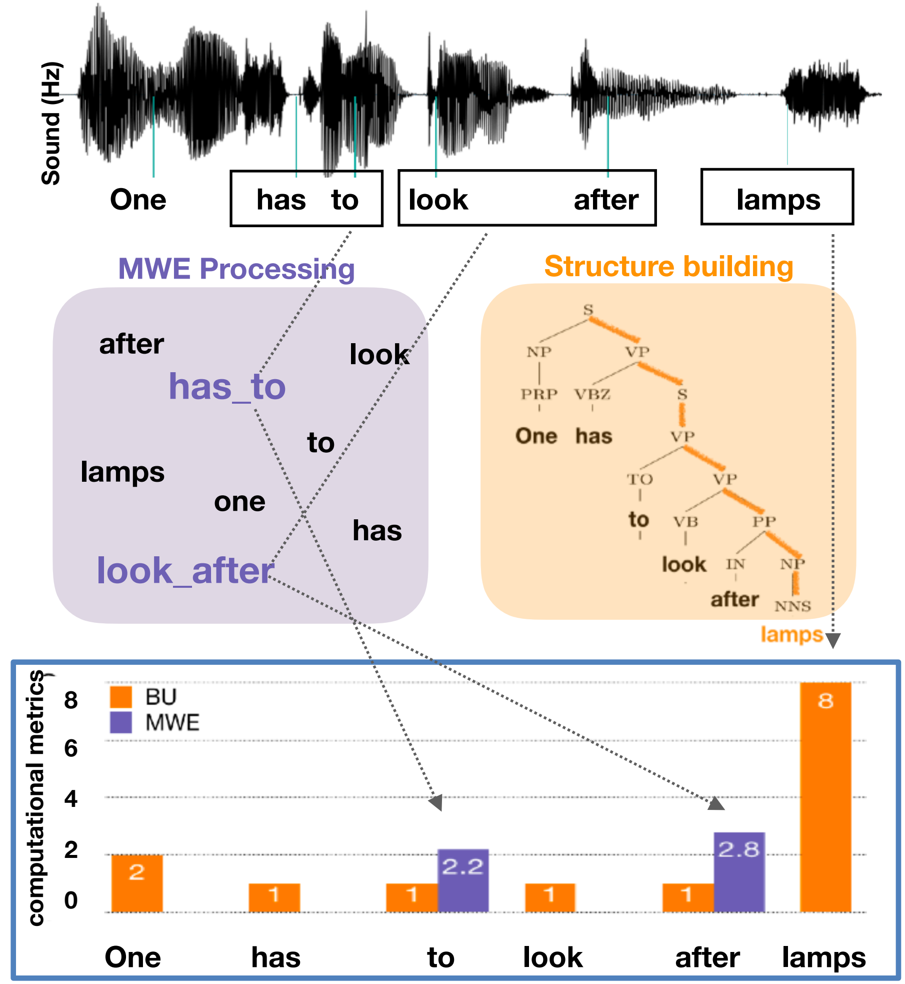


Scientific Questions and Hypotheses

Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.



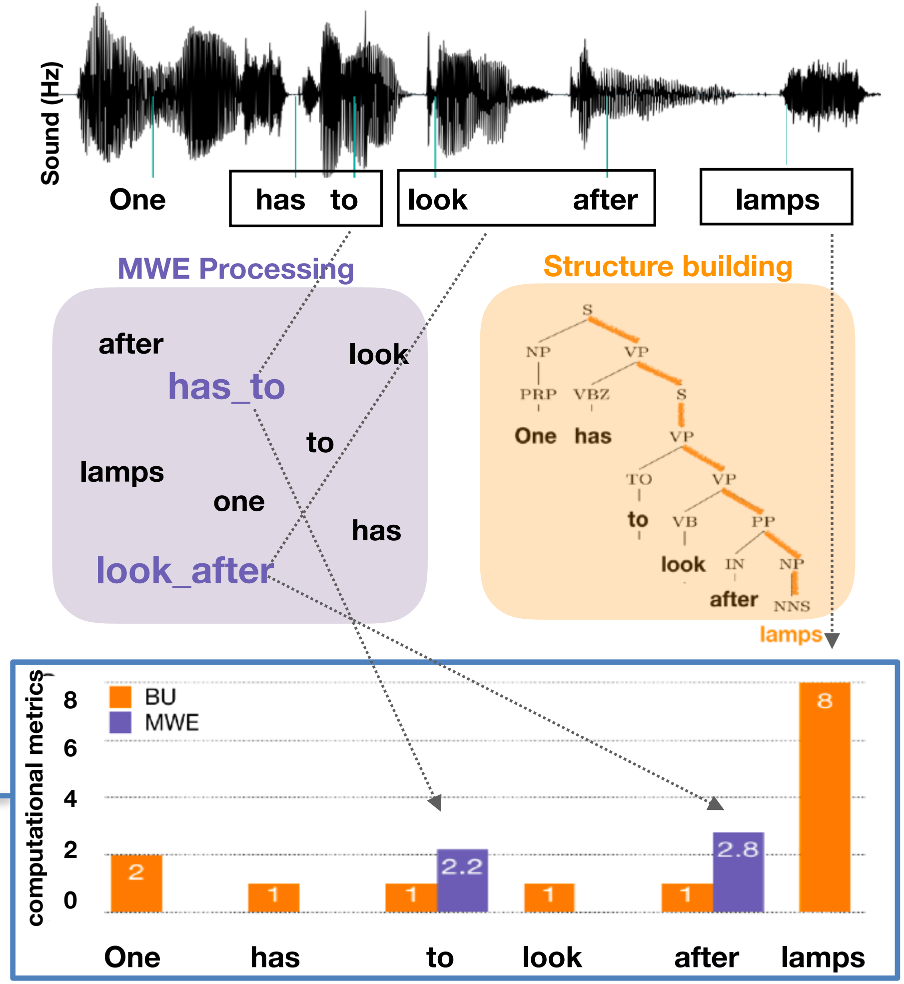
PMI a computational graded quantification identifying expressions likely to be processed as units, rather than built-up compositionally, BU tracks tree-building work needed in composed syntactic phrases.

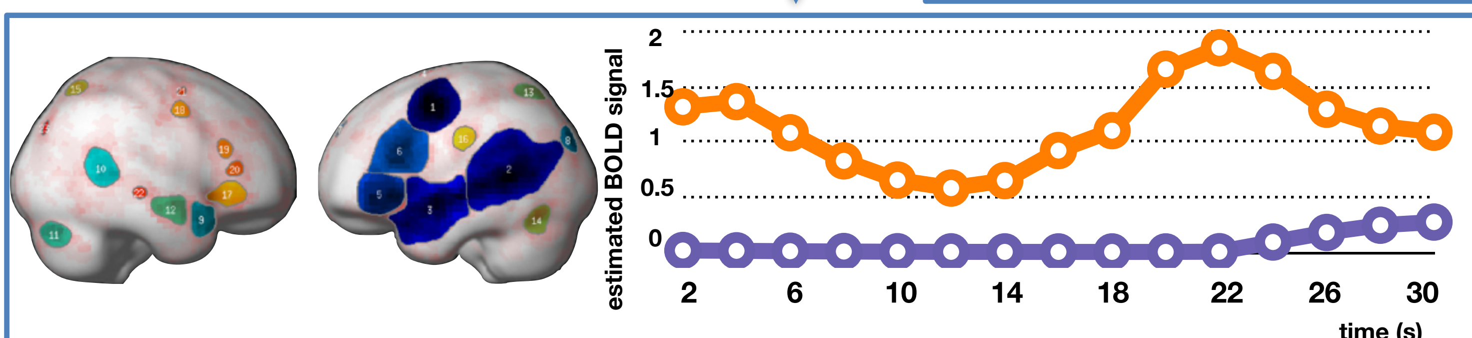
Scientific Questions and Hypotheses

Multi-word expressions

Frequently co-occurring word sequences, known as Multiword Expressions (MWEs) are likely to be processed differently by the language network.

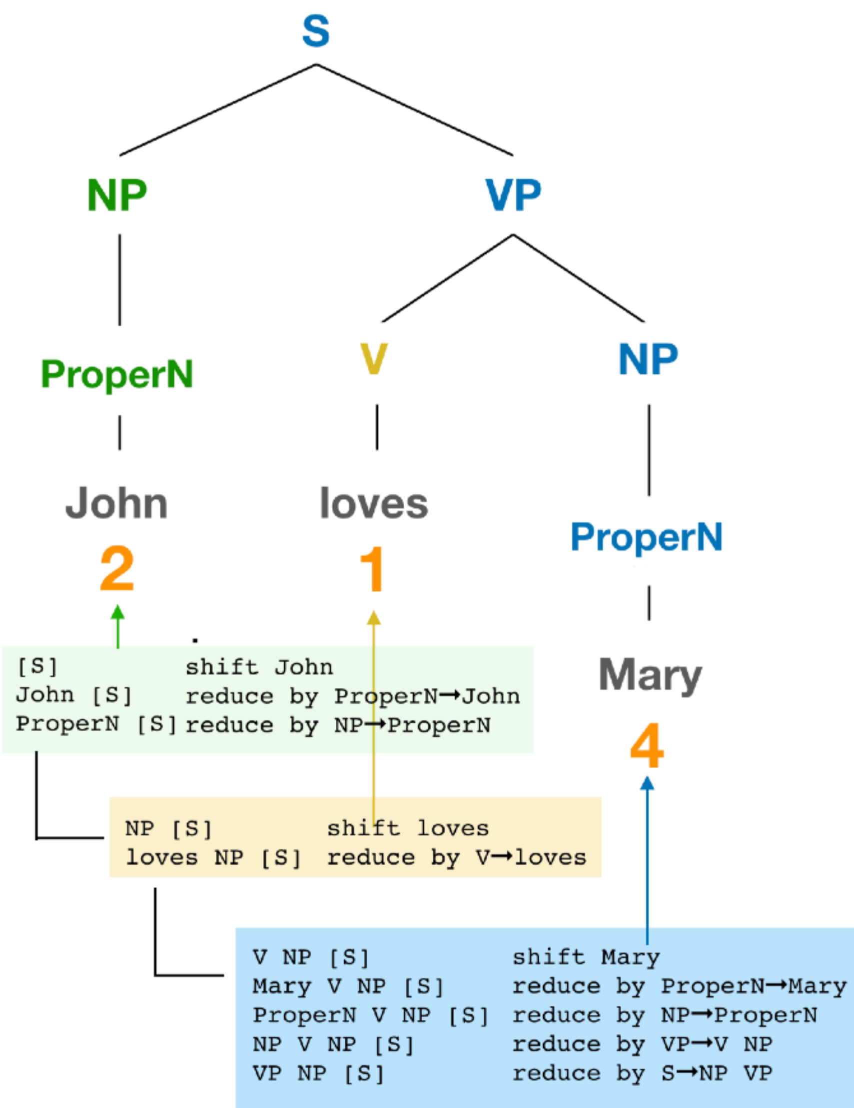
MWEs are a perfect testing ground to understand how expressions like ***break the ice***, ***boa constrictor***, ***see to it***, ***in spite of*** are processed in the brain.





PMI a computational graded quantification identifying expressions likely to be processed as units, rather than built-up compositionally, BU tracks tree-building work needed in composed syntactic phrases.

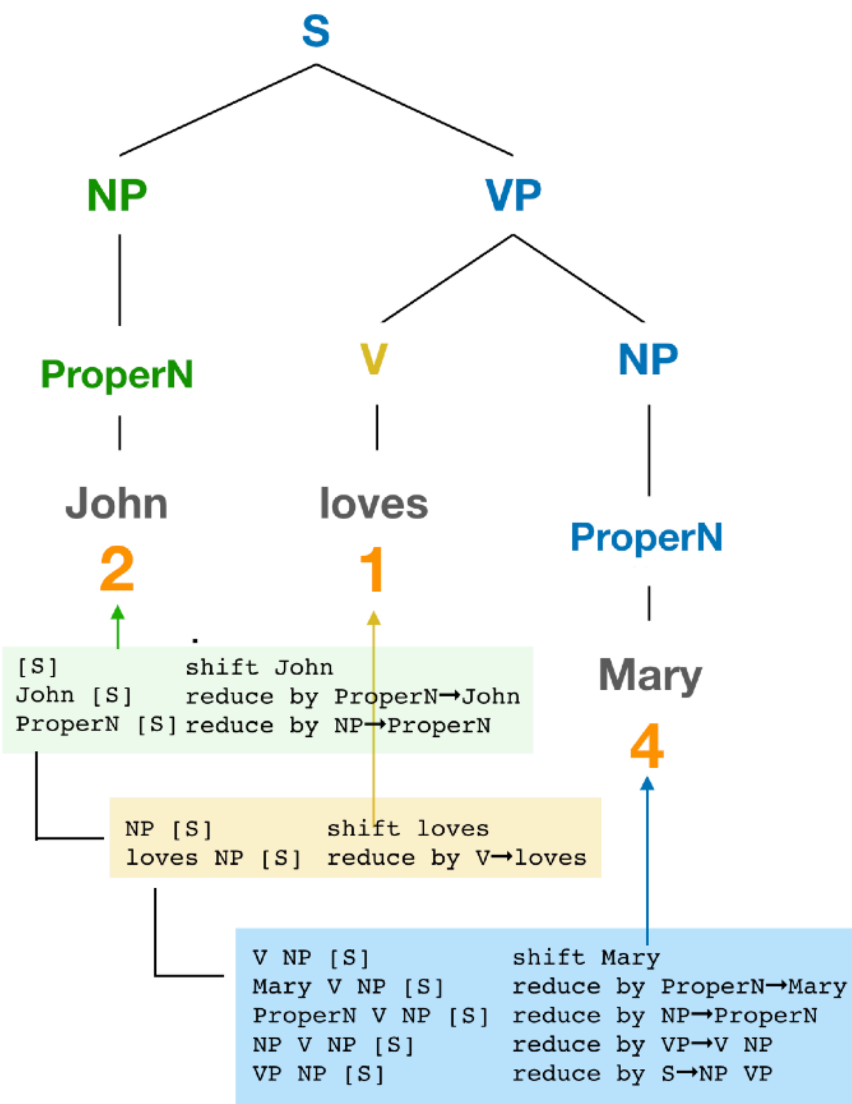
Investigating syntax in the brain through parsers



Bottom-Up Parser actions count

Hierarchical representation → **Parser actions**

Investigating syntax in the brain through parsers



Parser actions Bottom-Up

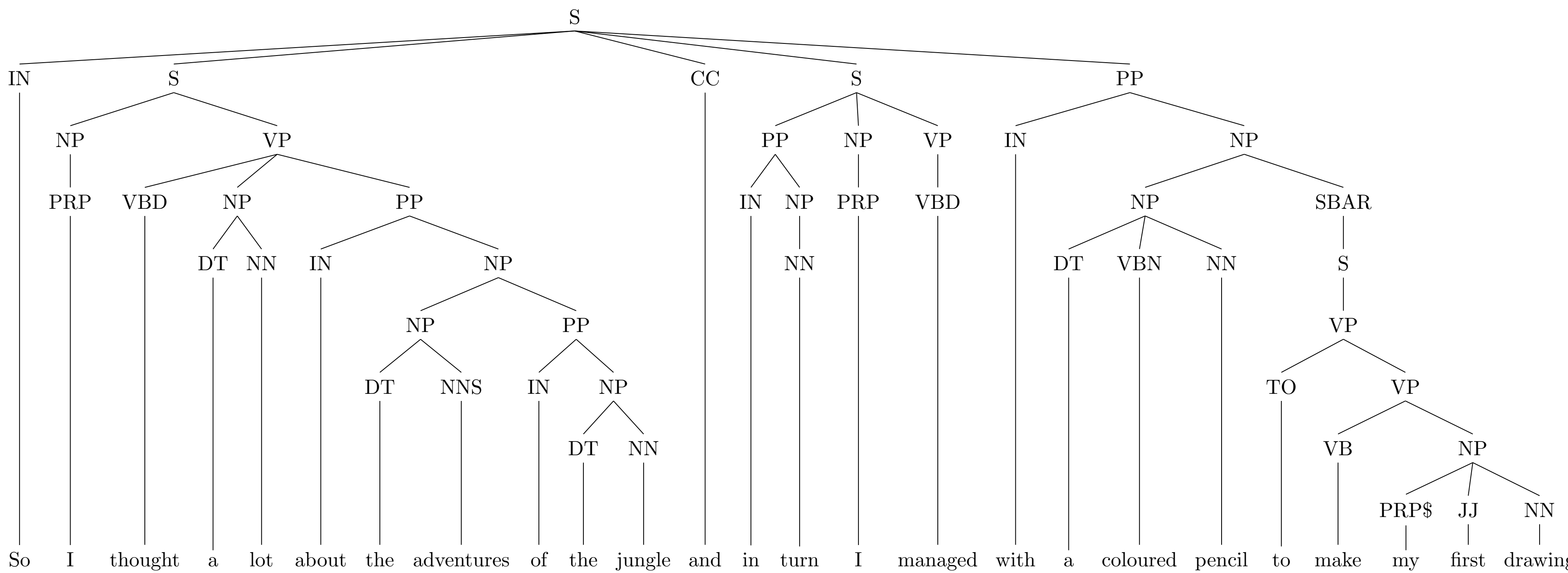
Hierarchical representation → **Parser actions**

Word-by-word measure of syntactic-structure building:

- **Can instantiate constituent-structure building the phrase/sentence.** as it builds and collects sub-parses towards the end of the phrase or sentence.
- **The rules of a grammar are applied at each incoming word**

Parser actions count

Sentence hierarchical representation + computational complexity metrics



Bottom-up parser action count

1	2	1	1	2	1	1	2	1	1	7	1	1	3	2	3	1	1	1	2	1	1	1	1	10

Number of REDUCE actions taken since last word

Association measure : Point-wise Mutual Information

PMI : A computational measure to link the degree of cohesiveness of MWEs

Computational graded **quantification identifying expressions likely to be processed as units, rather than built-up compositionally**

Association measure : Point-wise Mutual Information

PMI : A computational measure to link the degree of cohesiveness of MWEs

Computational graded quantification identifying expressions likely to be processed as units, rather than built-up compositionally

$$\text{PMI} = log_2\left(\frac{O}{E}\right)$$

where

$$O = \frac{count(\text{whole expression})}{\text{corpus size}}$$

and

$$E = \frac{\text{count}(w_1) * \text{count}(w_2) * \cdots * \text{count}(w_n)}{\text{corpus size}^n}$$

American english corpus : Coca 560 millions

Association measure : Point-wise Mutual Information

PMI : A computational measure to link the degree of cohesiveness of MWEs

$$\text{PMI} = log_2\left(\frac{O}{E}\right)$$

where

$$O = \frac{count(\text{whole expression})}{\text{corpus size}}$$

and

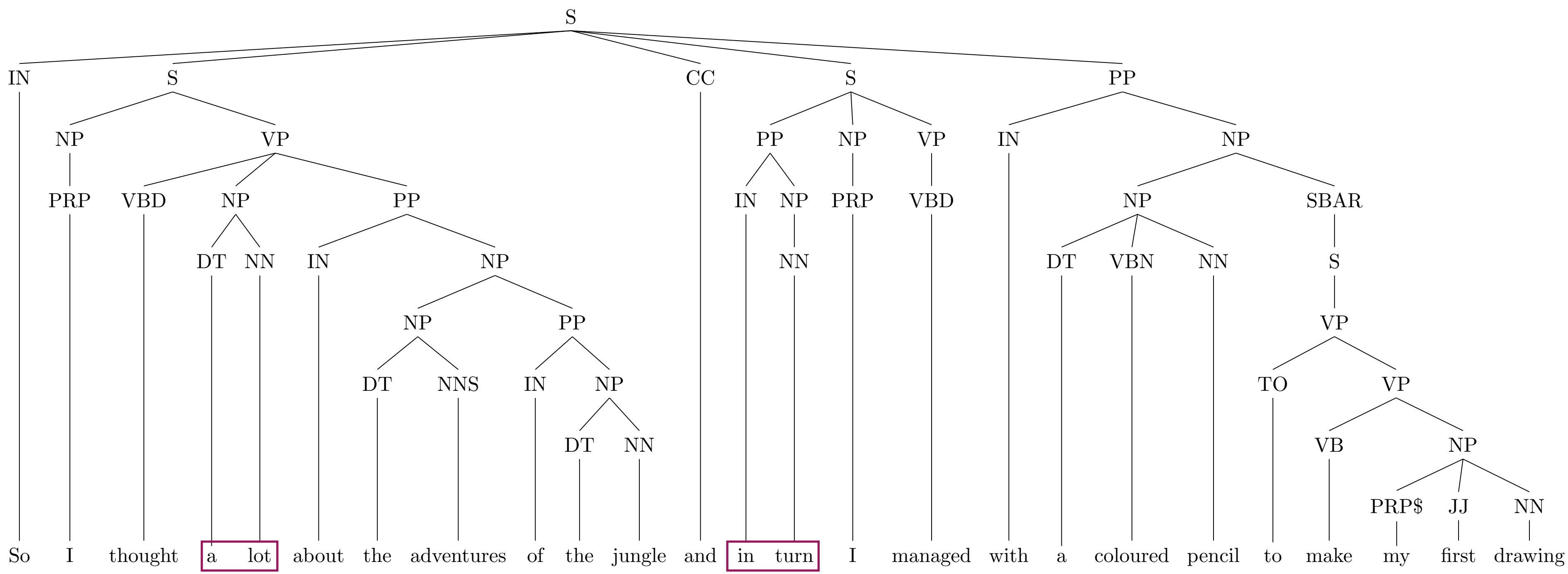
$$E = \frac{\text{count}(w_1) * \text{count}(w_2) * \cdots * \text{count}(w_n)}{\text{corpus size}^n}$$

PMI	multiword expression receiving this score
26.59474426	heart skipped a beat
23.79983038	have nothing to do with
21.25998782	forehead with a handkerchief
21.17721316	burst into tear
20.17480668	once upon a time
20.15121667	boa constrictor
18.85209561	peal of laughter
-2.336733827	be order
-2.493268369	do calculation
-2.721901963	be object
-2.982215241	be hundred
-3.152845604	a well
-3.501675488	drink anything
-3.635409951	have plan

American english corpus : Coca 560 millions

Parser actions count

Sentence hierarchical representation + computational complexity metrics



Bottom-up parser action count

1	2	1	1	2	1	1	2	1	1	7	1	1	3	2	3	1	1	1	2	1	1	1	1	10

MWE cohesion strength -> PMI

0	0	0	0	2.62	0	0	0	0	0	0	0	0	4.083	0	0	0	0	0	0	0	0	0	0	0

Identifying MWEs : English B

Computational toolkit ot identify MWEs

You

must

see

to

it

that

you

..

Computational toolkit ot identify MWEs

MWEs were identified using a statistical tagger (Al Saied et al. 2017). trained on Children's Book Test dataset.

You

must

see

to

it

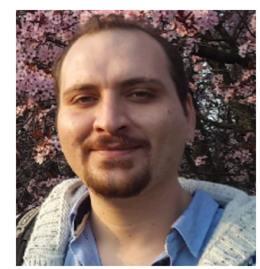
that

you

..

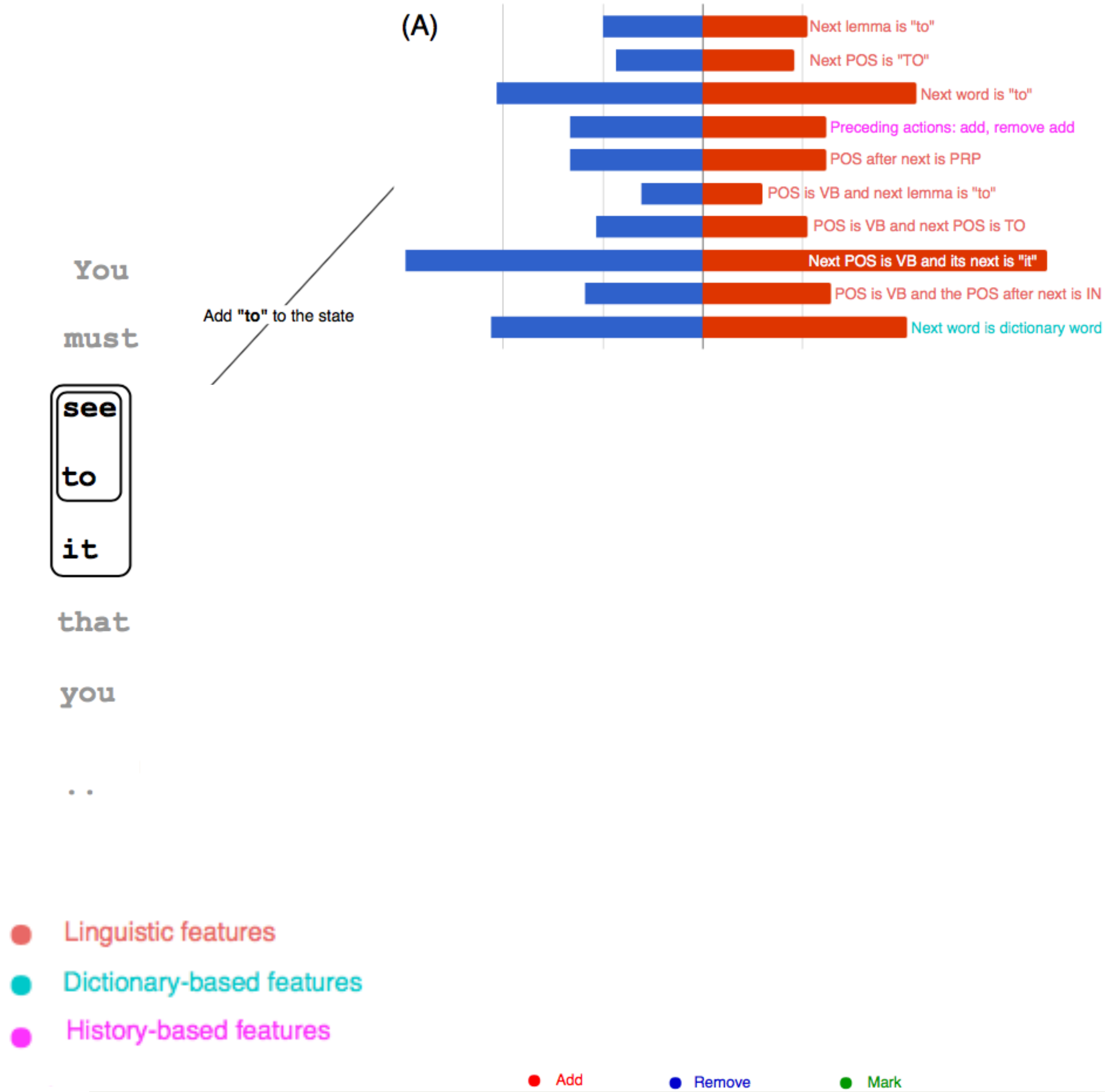
- Linguistic features
- Dictionary-based features
- History-based features

Mathieu Constant



Hazem Al-saied

Computational toolkit ot identify MWEs

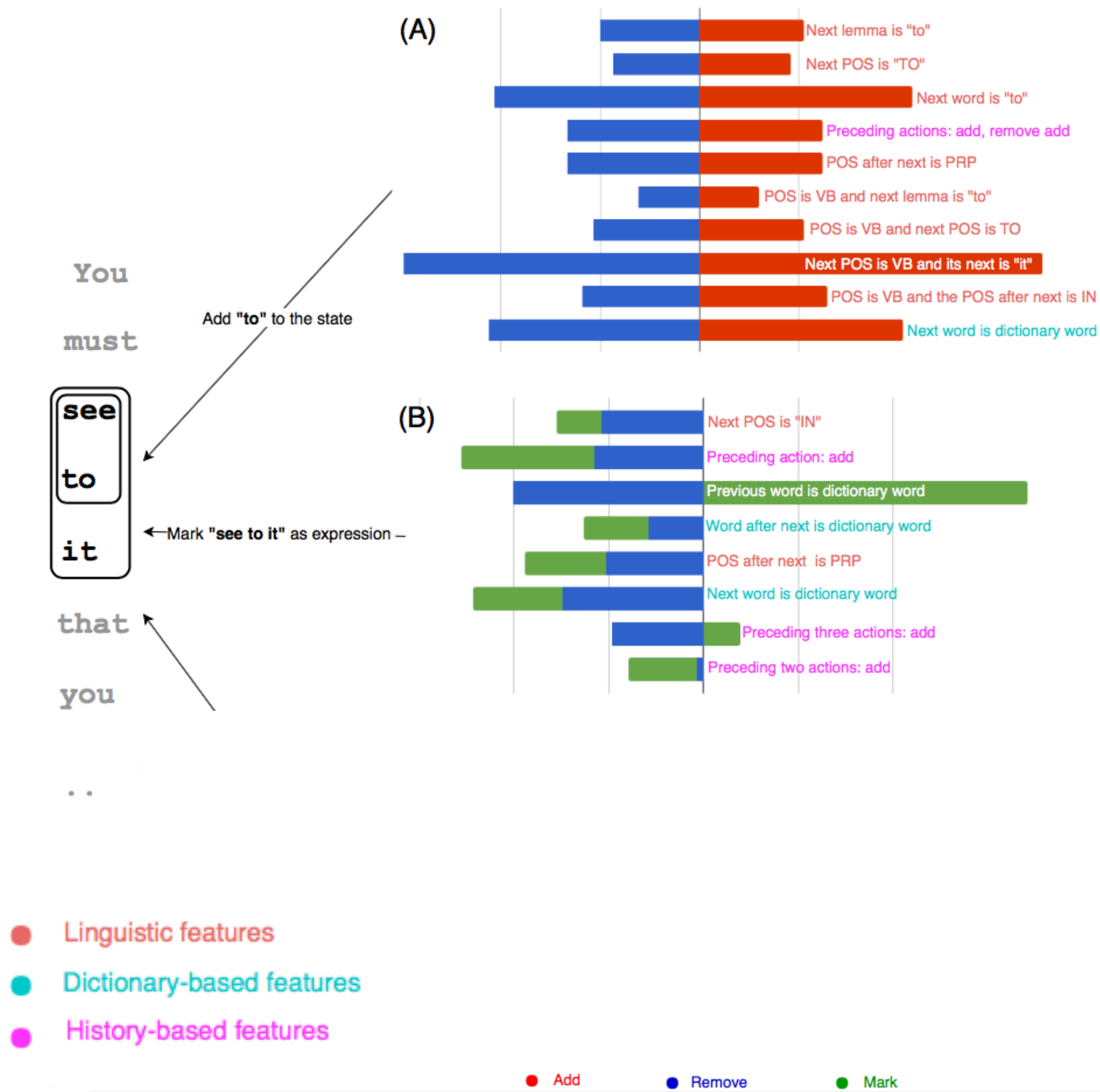


MWEs were identified using a statistical tagger (Al Saied et al. 2017). trained on Children's Book Test dataset.

Mathieu Constant

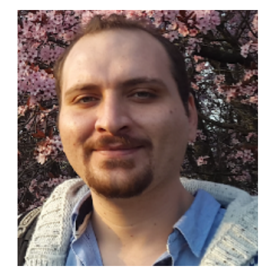
Hazem Al-saied

Computational toolkit ot identify MWEs



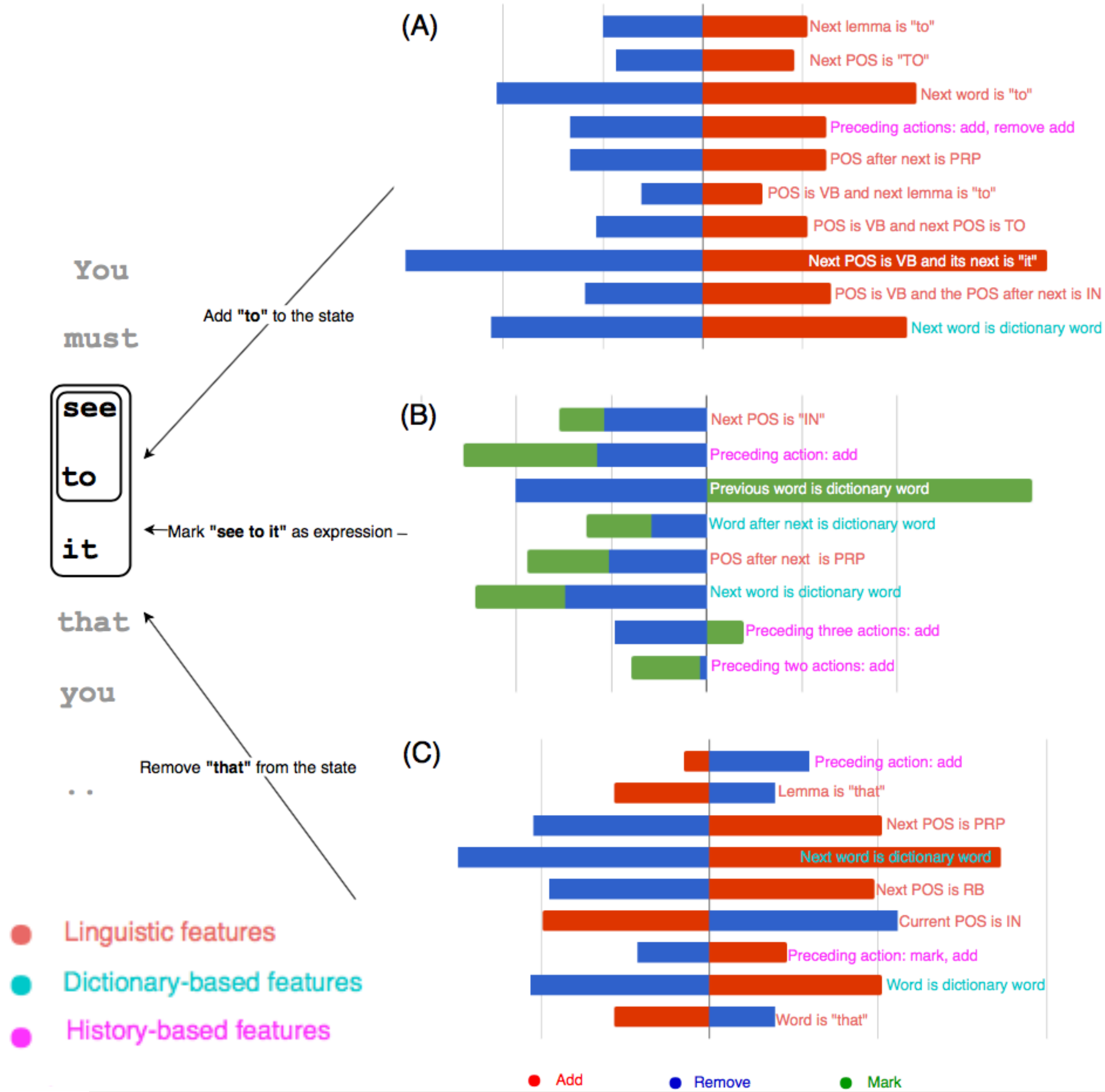
MWEs were identified using a statistical tagger (Al Saied et al. 2017). trained on Children's Book Test dataset.

Mathieu Constant



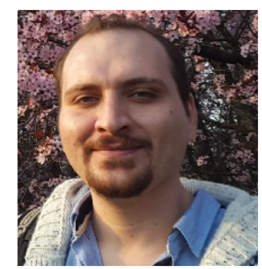
Hazem Al-saied

Computational toolkit to identify English MWEs



MWEs were identified using a statistical tagger (Al Saied et al. 2017). Trained on Children's Book Test dataset.

Mathieu Constant



Hazem Al-saied

MWE English

$w_t = X, t \in \{2, 1, 0, 1, 2\}$	$\&l_0 = L$
Lowercase form of $w_0 = W$	$\&l_0 = L$
Prefix of $w_0 = Pwith\|P\|<5$	$\&l_0 = L$
Suffix of $w_0 = Swith\|S\|<5$	$\&l_0 = L$
w_0 contains a hyphen	$\&l_0 = L$
w_0 contains a digit	$\&l_0 = L$
w_0 is capitalized	$\&l_0 = L$
w_0 is all in capital	$\&l_0 = L$
w_0 is capitalized and BOS	$\&l_0 = L$
w_0 is part of a multiword	$\&l_0 = L$
$w_i w_j = XY, (j, k) \in \{(1, 0), (0, 1), (1, 1)\}$	$\&l_0 = L$
$l_{-1} = L'$	$\&l_0 = L$

Table 1: Feature templates to detect MWEs

MWE English

$w_t = X, t \in \{2,1,0,1,2\}$	$\&l_0 = L$
Lowercase form of $w_0 = W$	$\&l_0 = L$
Prefix of $w_0 = Pwith\|P\|<5$	$\&l_0 = L$
Suffix of $w_0 = Swith\|S\|<5$	$\&l_0 = L$
w_0 contains a hyphen	$\&l_0 = L$
w_0 contains a digit	$\&l_0 = L$
w_0 is capitalized	$\&l_0 = L$
w_0 is all in capital	$\&l_0 = L$
w_0 is capitalized and BOS	$\&l_0 = L$
w_0 is part of a multiword	$\&l_0 = L$
$w_i w_j = XY, (j,k) \in \{(1,0),(0,1),(1,1)\}$	$\&l_0 = L$
$l_{-1} = L'$	$\&l_0 = L$

Table 1: Feature templates to detect MWEs

121 my$_{1}$ friend$_{2}$ broke$_{3}$ into$_{4}$ another$_{5}$ peal$_{6}$ of$_{7}$ laughter$_{8}$:$_{9}$ \`\`$_{10}$ where$_{11}$ do$_{12}$ you$_{13}$ think$_{14}$ he$_{15}$ 'd$_{16}$ go$_{17}$!$_{18}$ ''$_{19}$

122 \`\`$_{1}$ anywhere$_{2}$.$_{3}$

123 straight$_{1}$ ahead$_{2}$ ...$_{3}$ ''$_{4}$ then$_{5}$ the$_{6}$ little$_{7}$ prince$_{8}$ said$_{9}$ gravely$_{10}$:$_{11}$ \`\`$_{12}$ that$_{13}$ does$_{14}$ n't$_{15}$ matter$_{16}$;$_{17}$ where$_{18}$ i$_{19}$ live$_{20}$,$_{21}$ everything$_{22}$ is$_{23}$ so$_{24}$ small$_{25}$!$_{26}$ ''$_{27}$

124 and$_{1}$ perhaps$_{2}$ with$_{3}$ a$_{4}$ hint$_{5}$ of$_{6}$ sadness$_{7}$,$_{8}$ he$_{9}$ added$_{10}$:$_{11}$ \`\`$_{12}$ straight$_{13}$ ahead$_{14}$ you$_{15}$ ca$_{16}$ n't$_{17}$ go$_{18}$ far$_{19}$ ...$_{20}$ ''$_{21}$

125 i$_{1}$ thus$_{2}$ learned$_{3}$ a$_{4}$ second$_{5}$ very$_{6}$ important$_{7}$ thing$_{8}$:$_{9}$ that$_{10}$ his$_{11}$ home$_{12}$ planet$_{13}$ was$_{14}$ barely$_{15}$ bigger$_{16}$ than$_{17}$ a$_{18}$ house$_{19}$!$_{20}$

126 it$_{1}$ did$_{2}$ n't$_{3}$ surprise$_{4}$ me$_{5}$ much$_{6}$.$_{7}$

127 i$_{1}$ knew$_{2}$ that$_{3}$,$_{4}$ apart$_{5}$ from$_{6}$ the$_{7}$ large$_{8}$ planets$_{9}$ like$_{10}$ the$_{11}$ earth$_{12}$,$_{13}$ jupiter$_{14}$,$_{15}$ mars$_{16}$,$_{17}$ and$_{18}$ venus$_{19}$,$_{20}$ which$_{21}$ have$_{22}$ been$_{23}$ given$_{24}$ names$_{25}$,$_{26}$ there$_{27}$ are$_{28}$ hundreds$_{29}$ of$_{30}$ others$_{31}$ that$_{32}$ are$_{33}$ sometimes$_{34}$ so$_{35}$ small$_{36}$ that$_{37}$ one$_{38}$ has$_{39}$ great$_{40}$ difficulty$_{41}$ in$_{42}$ spotting$_{43}$ them$_{44}$

MWE English

The audio stimulus was Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet.

Within this text, 1,274 MWEs were identified using a CRF tagger. This tagger was trained on examples from the English Universal Dependency treebank, in combination with external lexicons as suggested by Constant and Tellier (2012). The tagger used feature templates, as seen in Table 1 below, where w_t stands for the token at the relative position t from the current token and l_t is the label at the relative position t. The external lexicons included the Unitex lexicon (Paumier et al., 2009), SAID corpus (Kuiper et al., 2003), Cambridge International Dictionary of Idioms (White, 1998), and Dictionary of American Idioms (Makkai et al., 1995).

Matthieu Constant and Isabelle Tellier. 2012. Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In *8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 646–650.

MWE English

The audio stimulus was Antoine de Saint-Exupéry's *The Little Prince*, translated by David Wilkinson and read by Nadine Eckert-Boulet.

MWE Category	Occurrence
Verb + Participle	145
Verb + Noun	37
Adj + Noun	285
Det + Noun	712
(Verb) + Noun + Prep + Noun	24
N-N Compounds	71

Table 2: MWE Attestation Rates

Within this text, 1,274 MWEs were identified using a CRF tagger. This tagger was trained on examples from the English Universal Dependency treebank, in combination with external lexicons as suggested by Constant and Tellier (2012). The tagger used feature templates, as seen in Table 1 below, where w_t stands for the token at the relative position t from the current token and l_t is the label at the relative position t. The external lexicons included the Unitex lexicon (Paumier et al., 2009), SAID corpus (Kuiper et al., 2003), Cambridge International Dictionary of Idioms (White, 1998), and Dictionary of American Idioms (Makkai et al., 1995).

Matthieu Constant and Isabelle Tellier. 2012. Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In *8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 646–650.

Identifying MWEs : French B

Automatized identification of French MWEs

Extracting MWE FRMG *patterns* matching Wiktionary entries

Verifying patterns in a second corpus

Filtering results

> to the average PMI

> to the average occurence

Euro-Parliament Corpus - 41,5 millions
both written and oral style

Wikisource Corpus - 64 millions
Narrative written style

Calculating
PMIs scores
&
Average occurence
in

French euro-parliament
+ French wikisource
+ French wikipedia
(179 millions)

Yoann Dupont

Eric de la Clergerie

Automatized identification of French MWEs

Extracting MWE FRMG *patterns* matching Wiktionary entries

Verifying patterns in a second corpus

Filtering results

> to the average PMI

> to the average occurence

Euro-Parliament Corpus - 41,5 millions
both written and oral style

Wikisource Corpus - 64 millions
Narrative written style

Calculating PMIs scores & Average occurence in
French euro-parliament + French wikisource + French wikipedia (179 millions)

Le Petit Prince Audio

Hand-picked selection

1300

a) point de sa chute
b) point de chute

Yoann Dupont

Eric de la Clergerie

MWE Patterns : French

-> Ordre des mots des adjectifs antéposé post-posés :

- donner une fausse idée
- pas m' étonner beaucoup
- ne éprouver *plus* le besoin

-> Avantage d'avoir une représentation en dépendances :

MWE longue

- entourer le cou de son bras

MWE avec inclusion

- entrer *à son tour* dans la danse

Figures de style imagées non-incluses

- s'enroula autour de sa cheville, comme un bracelet d'or
- entasser l'humanité sur un îlot

MWE patterns English vs. French

English

MWE Category	Occurrence
Verb + Participle	145
Verb + Noun	37
Adj + Noun	285
Det + Noun	712
(Verb) + Noun + Prep + Noun	24
N-N Compounds	71

Table 2: MWE Attestation Rates

French

MWE Category	Occurrence
Verbale	631
Nominales	380
Adverbiales	32
Prépositionnelles	305

Table 3: MWE Attestation Rates

Identification of French MWEs : patterns

Adverbial

tout doucement
bien loin
peut-être bien que
tout à fait
un peu juste
oui ou non
à peine plus

Nominal

oeil clos
passant ordinaire
pas de course
couleur de miel
drôle de bête
éclat de rire
économie de temps
geste de lassitude
mèche de cheveu
messe de minuit
peine de mort
mouvement de regret
poupée de chiffon
source de malentendu

Prepositional

au fond de son coeur
à tout hasard
ni faim ni soif
sur terre
comme un fontaine dans le désert
contre tout espérance
en larme
faute de patience

Verbal

clr habiller à le européen
clr voir important comme
écraser son nez contre le vitre
ébaucher un sourire
être bien obliger
lever le oeil vers le ciel
ne clr avancer pas à grand-chose
ne manger pas de pain
apaiser le soif
aimer les chiffres
avoir mouiller le tempe
boire le dernier goutte
ce ne être pas mon faute
cld habiller le coeur
clr enfoncer dans une rêverie
parler toujours le premier

Stability of PMI scores

C

Preliminary steps to calculate PMIs - **French**

1. Building of a corpus of a comparable size to the COCA coprus.
2. Building a corpus of children books comparable to the CBT
3. Dependency parsing to capture more open MWEs

French Children books Corpus

Children's books Test Corpus

	CBT English	→	CBT Français
Size	108 books 6 millions		Wikisource - Gutenberg 6 millions
Register	livres et histoires pour la jeunesse	→	Littérature jeunesse et des classiques contenants des dialogues

Go through Project Gutenberg to find children's books, then stripping out the Project Gutenberg headers (which is sadly nontrivial). They have a lot of public domain works already transcribed in .txt form.

Children's book dataset taken as reference : fb.ai/babi / http://arxiv.org/abs/1511.02301

French *coca-style* Corpus

COCA corpus (Davies, 2008)
Corpus of Contemporary American English

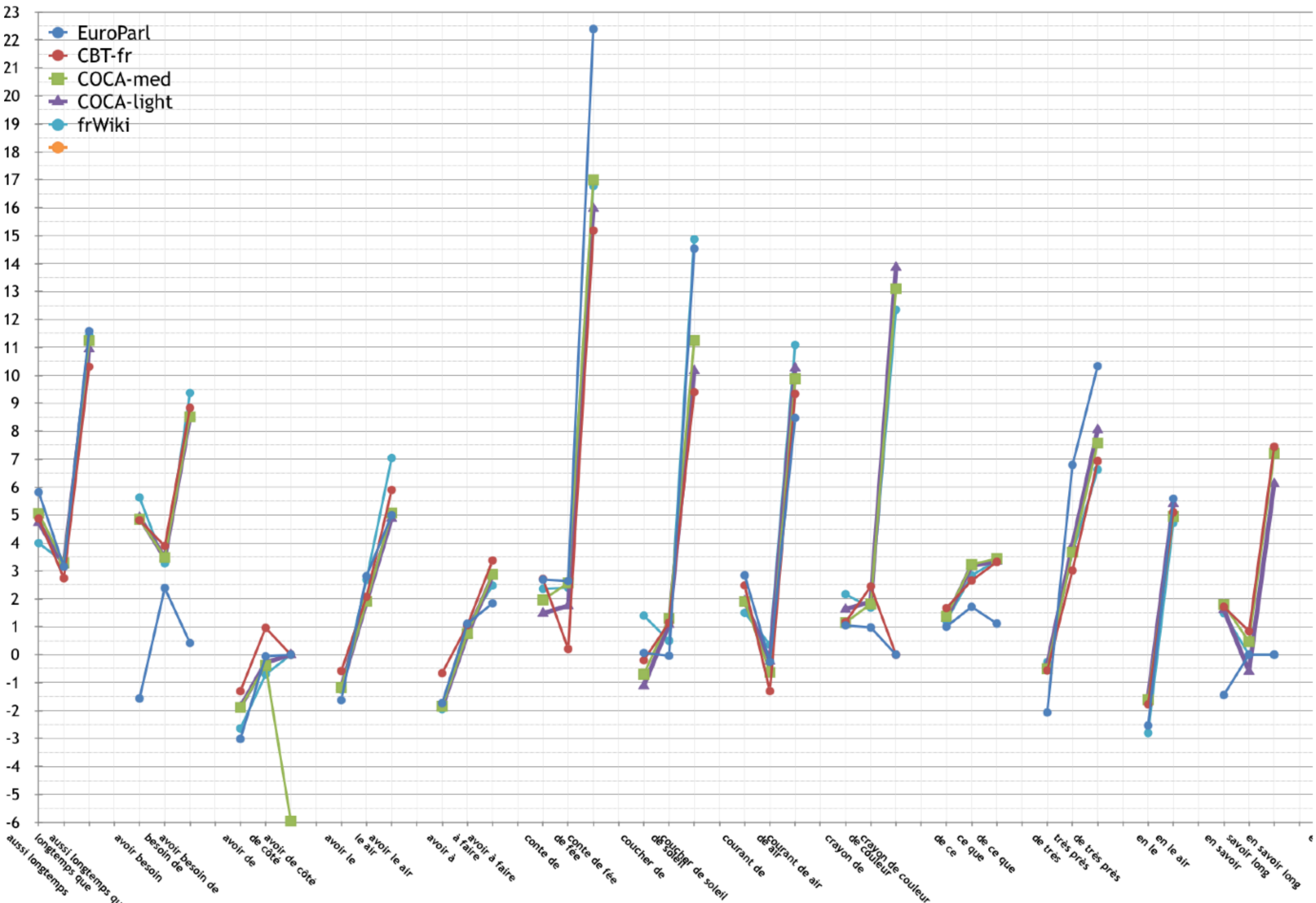
	COCA Américain →	***COCA-fr* Français**
Size	560 millions 20 million words each year 1990-2017	500 millions
Register	spoken, fiction, popular magazines, newspapers, and academic texts →	**spoken, fiction, popular magazines, newspapers, and academic texts**

Leverage on already published open access corpora + stripping out Whikisource headers (which is sadly nontrivial) and constitute an agglomerated corpus of public domain works transcribed in .txt form.

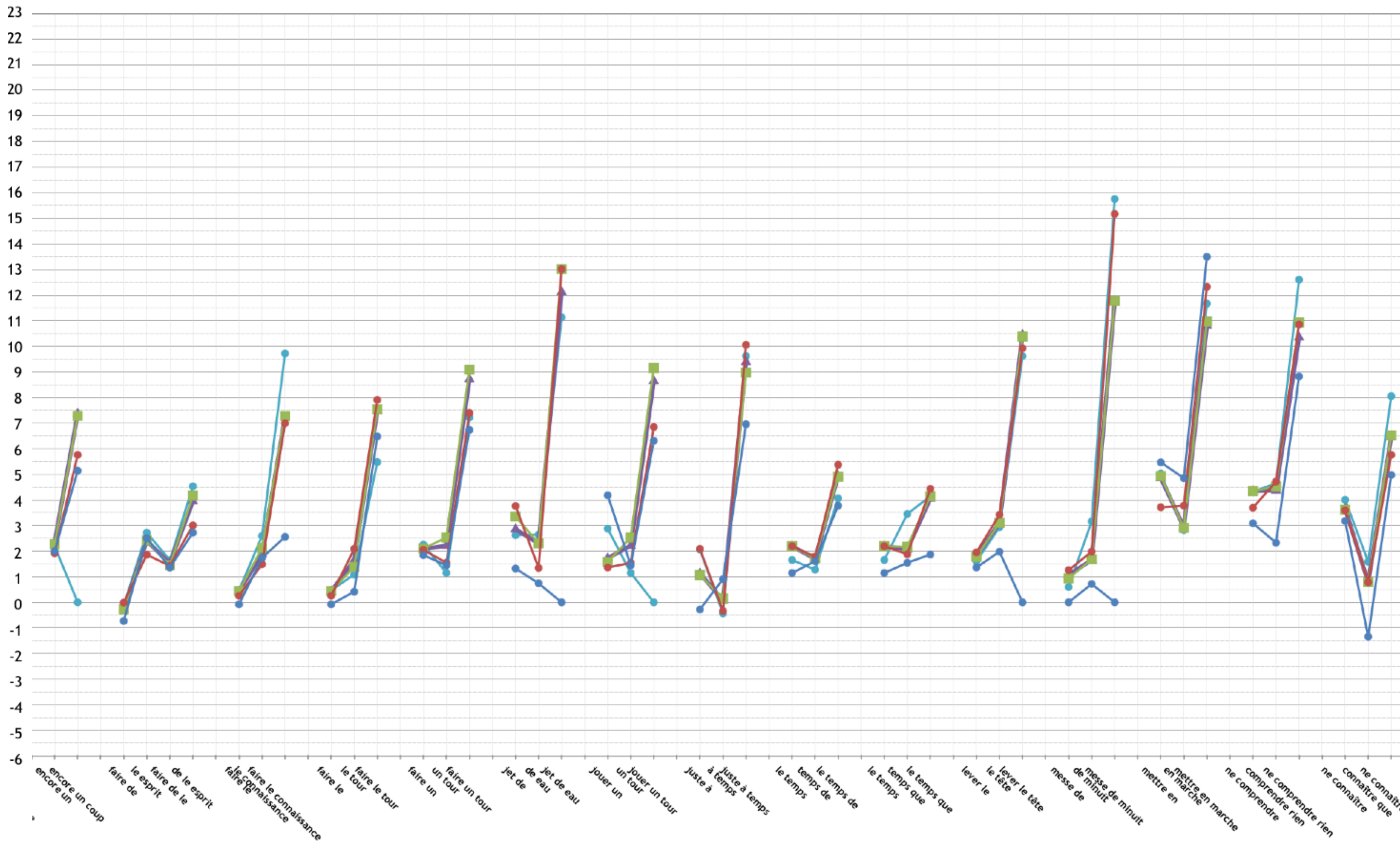
Stability of PMI scores across corpora

	41,5 m	6 m	200 m	6 m	180 m
MWE	EuroParl	CBT-fr	COCA-med	COCA-light	frWiki
aussi longtemps	5,81581737575408	4,8702808038489716	5,055847397536948	4,728799208717715	3,9948403736050615
longtemps que	3,1528370542282764	2,7375650839441907	3,284503316083517	3,2167065147065794	3,3056193662774658
aussi longtemps que	11,582403946175281	10,304810567804275	11,24114681240891	10,946023839802988	11,43912108914338
avoir besoin	-1,5653279015342294	4,810792700043313	4,846303301719884	4,917171252847821	5,62495966318971
besoin de	2,385653111797074	3,8912354312130146	3,480141804686193	3,464294884590617	3,2684480955926682
avoir besoin de	0,41922744149313007	8,83515248595403	8,51335832420923	8,529510845290394	9,369162904962081
avoir de	-3,014932185746996	-1,3062217147456807	-1,8859488402796416	-1,873665419400638	-2,638976601880579
de côté	-0,05682058476818016	0,9612101368952901	-0,3837295717221365	-0,2814271940761609	-0,7076132324797088
avoir de côté	0,0	0,0	-5,954933684547455	0,0	0,0
avoir le	-1,6303065946471786	-0,5815478797733671	-1,173460098390195	-1,163487896061267	-1,2394297883619239
le air	2,8099216956927653	2,074618620707513	1,9077948951325125	1,8969314605246372	2,6820639398119623
avoir le air	4,992966111872359	5,897051420996743	5,073380401864775	4,886973030767639	7,037703230107961
avoir à	-1,7350829567286645	-0,6627455445870685	-1,8430753541196299	-1,878871745756357	-1,954660235371303
à faire	1,1088580857233334	1,0518484832274129	0,7557687077065567	0,7169916701136999	1,0219498514795022
avoir à faire	1,837153915096479	3,3722509118483175	2,877626600974062	2,8227609474102024	2,4751585478637095
conte de	2,6886895447261927	2,710784473342912	1,9622193435352233	1,4799411380091538	2,3576662284620054
de fée	2,6346128719704	0,20461612079513486	2,569506026167946	1,7541359895801698	2,395305080465858
conte de fée	22,394410864432988	15,181917201459925	16,992310796960687	15,973423778857041	16,77398928957456

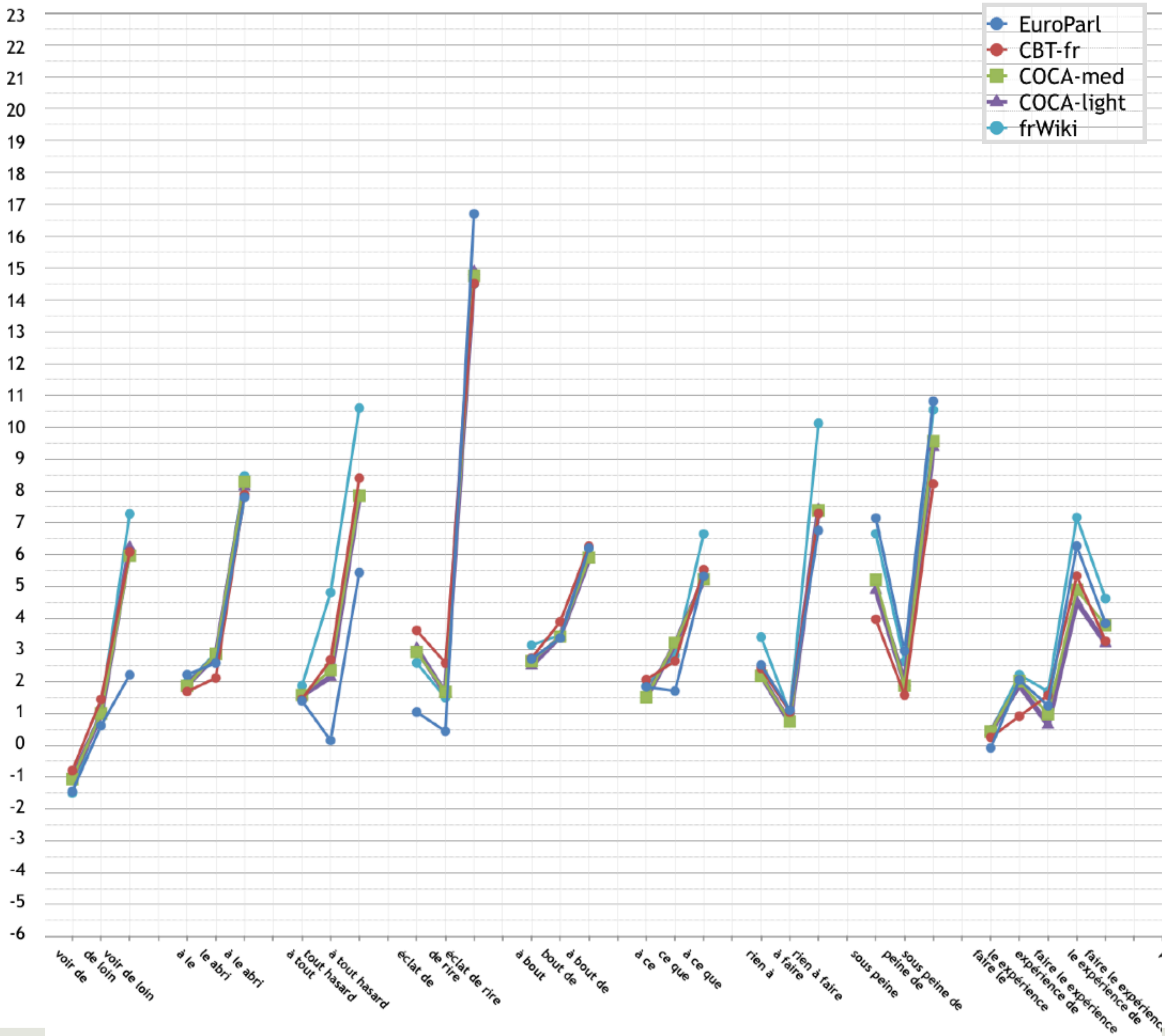
Stability of PMI measures



Stability of PMI measures



Stability of PMI measures

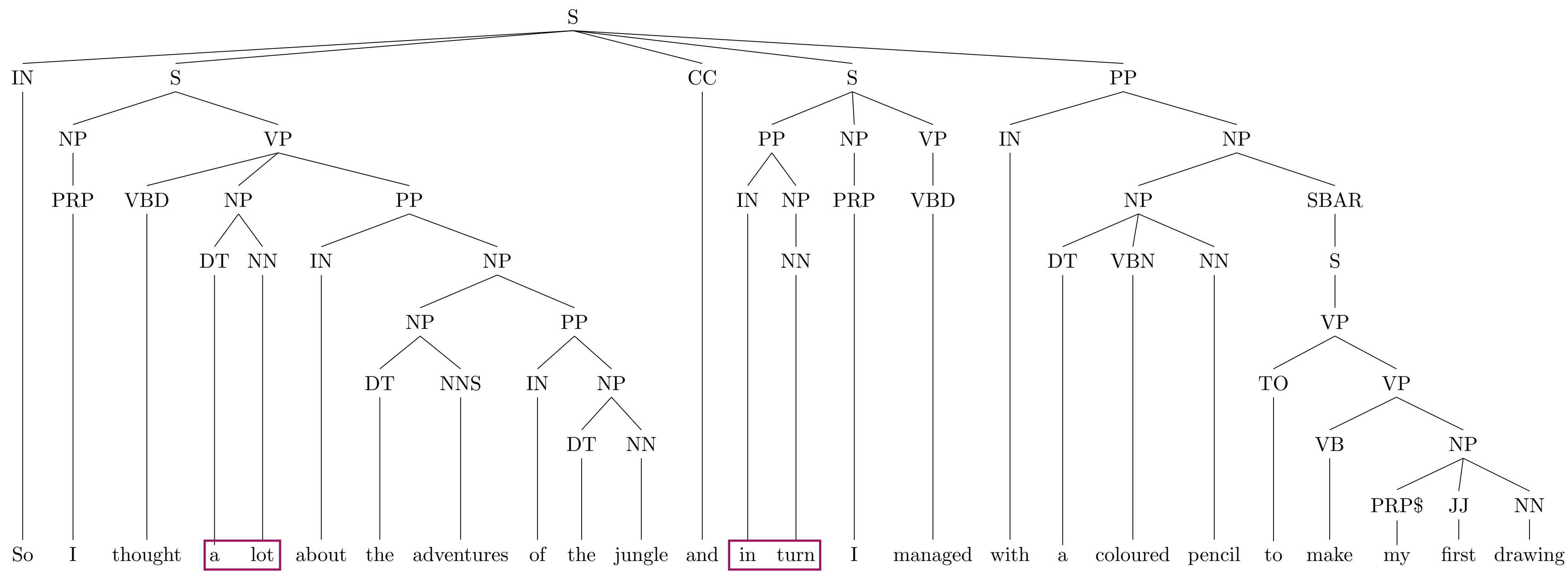


fMRI results - English

D

Parser actions count

Sentence hierarchical representation + computational complexity metrics



Bottom-up parser action count

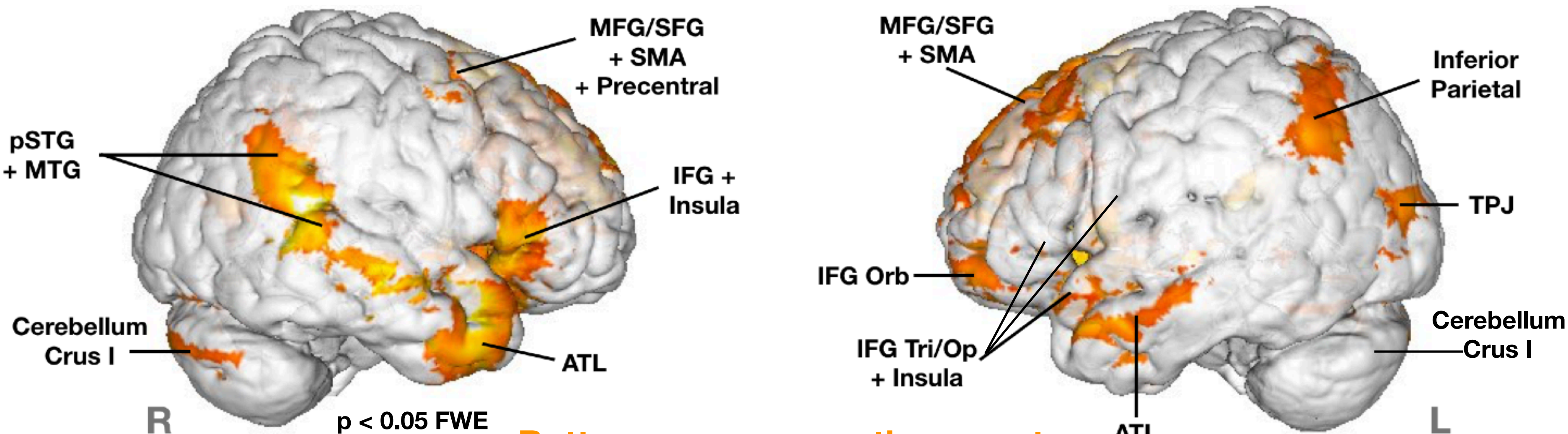
1	2	1	1	2	1	1	2	1	1	7	1	1	3	2	3	1	1	1	2	1	1	1	1	10

MWE cohesion strength -> PMI

0	0	0	0	2.62	0	0	0	0	0	0	0	0	4.083	0	0	0	0	0	0	0	0	0	0	0

PMI a computational graded quantification identifying expressions likely to be processed as units, rather than built-up compositionally, BU tracks tree-building work needed in composed syntactic phrases.

Bottom- up parser action count - **English**

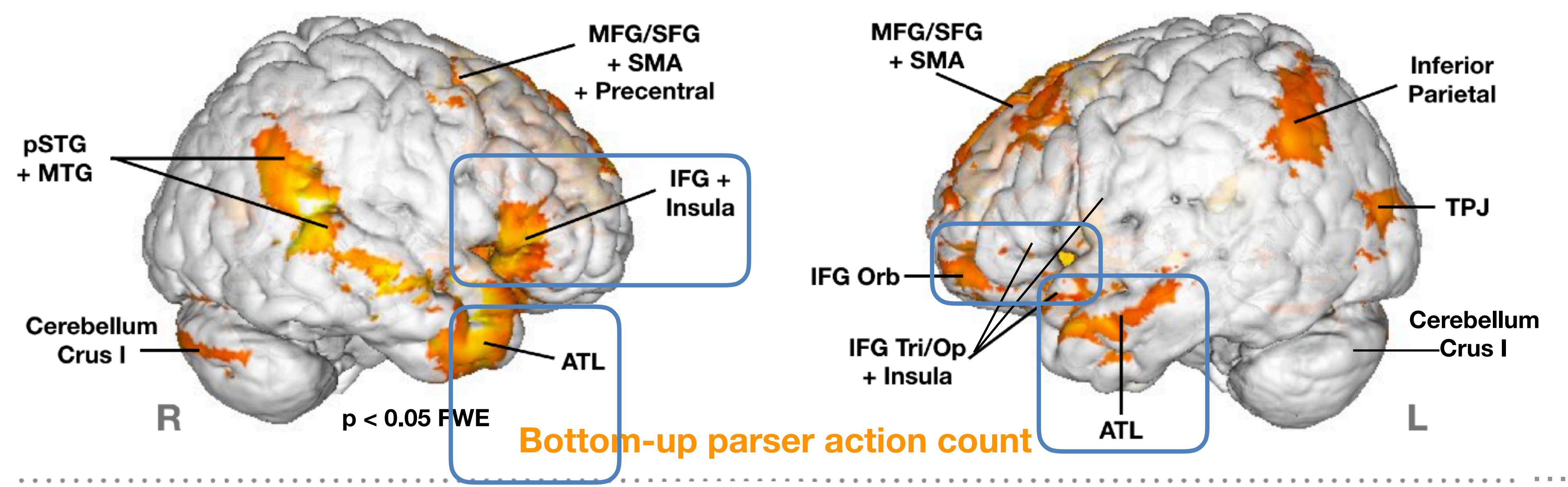


Bottom-up parser action count

→ **Bilateral network involving IFG and ATL**

Analysis of MWEs and parser action counts in naturalistic spoken story comprehension supports a dissociation between Temporal and Parietal brain structures and anterior Frontal regions such as IFG and ATL, as respectively sub-serving the retrieval of memorized expressions and structure-building processes.

Bottom- up parser action count - **English**



→ **Bilateral network involving IFG and ATL**

Analysis of MWEs and parser action counts in naturalistic spoken story comprehension supports a dissociation between Temporal and Parietal brain structures and anterior Frontal regions such as IFG and ATL, as respectively sub-serving the retrieval of memorized expressions and structure-building processes.

Positive and negative correlation with PMI - **English**

Increasing MWE cohesion strength (PMI)

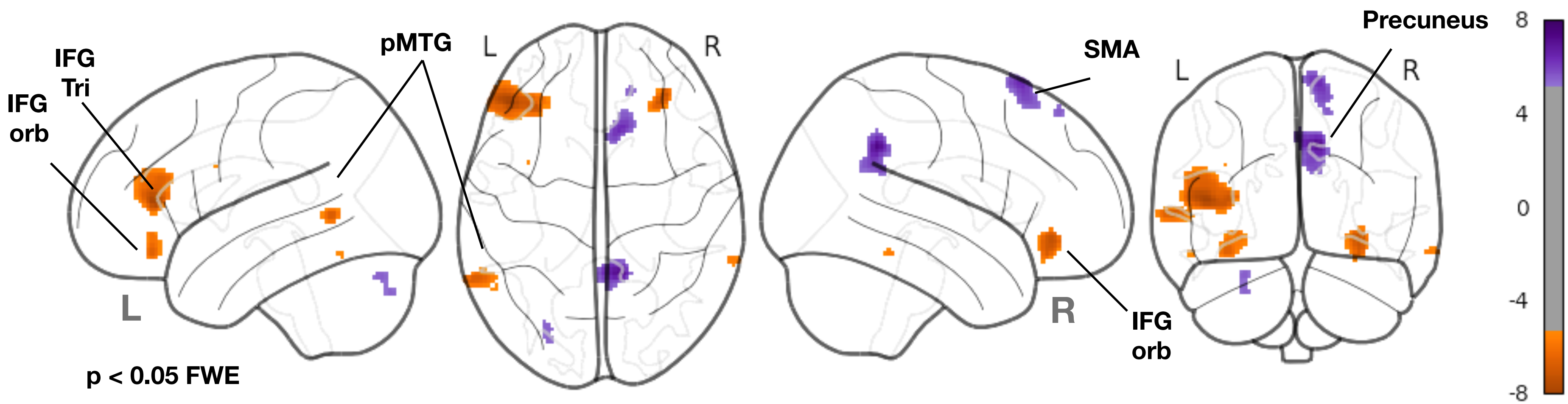
Decreasing MWE cohesion strength (PMI)

The results show an overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe. Highly cohesive MWEs implicate the Precuneus and the SMA, suggesting that only truly lexicalized linguistic expressions rely on these areas rather than traditional frontal and temporal nodes of the language network.

-> PMI as a proxy of lexical cohesiveness

Positive and negative correlation with PMI - **English**

Increasing MWE cohesion strength (PMI) **Decreasing MWE cohesion strength (PMI)**

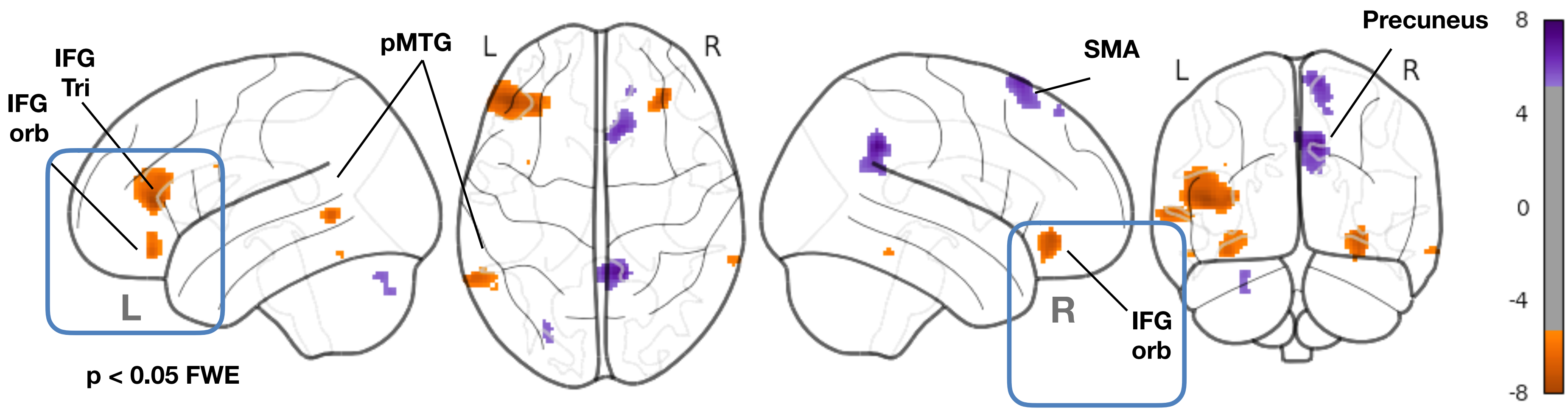


The results show an overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe. Highly cohesive MWEs implicate the Precuneus and the SMA, suggesting that only truly lexicalized linguistic expressions rely on these areas rather than traditional frontal and temporal nodes of the language network.

-> PMI as a proxy of lexical cohesiveness

Positive and negative correlation with PMI - **English**

Increasing MWE cohesion strength (PMI) **Decreasing MWE cohesion strength (PMI)**



The results show an overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe. Highly cohesive MWEs implicate the Precuneus and the SMA, suggesting that only truly lexicalized linguistic expressions rely on these areas rather than traditional frontal and temporal nodes of the language network.

-> PMI as a proxy of lexical cohesiveness

fMRI Results Summary

1 - PMI : word-by-word computational measure of lexical cohesiveness

—> cognitively plausible computational measure of the balance between *compositionality* versus *cohesiveness* in MWEs

Our study is showing that this association measure in MWEs produces a neuro-cognitive effect during naturalistic story listening.

2 - Effect of less cohesive MWEs and phase-structure building effect (BU)

We observe an overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe.

The results show an overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe. Highly cohesive MWEs implicate the Precuneus and the SMA, suggesting that only truly lexicalized linguistic expressions rely on these areas rather than traditional frontal and temporal nodes of the language network.

Next steps in French

1 - PMI : word-by-word computational measure of lexical cohesiveness <u>in French</u>

—> Confirm that PMI is a cognitively plausible computational measure of the balance between compositionality and cohesiveness in MWEs

- Compare French and English on the different degrees of compositionality measured with PMI scores
- Compare French and English in terms of morphology of the identified MWEs : Nominal versus Verbal patterns

2 - Effect of less cohesive MWEs and phase-structure building effect (BU) <u>in French</u>

- Replicate the overlap between the significant effect for decreasing MWE cohesiveness and Bottom-up parser action count in left IFG and posterior temporal lobe.

- Leverage on the open MWEs identified thanks to dependency parsing, and observe if different processes/activation patterns are elicited by **open versus contiguous MWEs**

Thanks for your attention