Expressions polylexicales dans la linguistique computationnelle: on n'est pas sorti de l'auberge

Agata Savary

Université Paris-Saclay, France

Université de Varsovie 18 mai 2022

(exposés similaires ont été donnés le 18 janvier 2019 et le 16 avril 2020)

Expressions polylexicales

Qu'y a-t-il 'quel est le problème' de spécial avec les expressions *mises en exergue* 'mises en évidence'?

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au rédacteur-en-chef, Monsieur **Jean-Marc Petit**.*

Expressions polylexicales

Définition

Combinaisons de **plusieurs mots** qui possèdent des **propriétés irrégulières** au niveau du lexique, de la grammaire, de la sémantique, etc.

Exemples de propriétés irrégulières

- **Sémantique non-compositionnelles**: le sens global n'est pas déductible de manière régulière à partir des sens des composants, et des liens syntaxiques qui les relient.
  - ***couper l'herbe sous le pied*** *de quelqu'un* 'empêcher quelqu'un de réussir'
  - ***s'en prendre*** *à quelqu'un* 'prendre quelqu'un pour cible, lui attribuer un faute'
- Déterminant zéro
  - ***avoir besoin***, **avoir un besoin*, ***avoir*** *un* ***besoin*** *important*
- Figement lexical:
  - ***rédacteur-en-chef*** , **journaliste-en-chef* , **rédacteur-en-leader*
- Figement syntaxique:
  - ***Jean-Marc Petit***, **Marc-Jean Petit*

Figement

Une EP est moins flexible (variable) qu'une construction régulière de la même structure syntaxique.

Construction régulière	**Expression polylexicale**	**Propriété**
livre bleu $\approx$[1] *livre cyan* $\approx$ *bouquin bleu*	***cordon bleu*** vs. *#cordon cyan* vs. *#corde bleue*	Figement lexical
manger des salades $\approx$ *manger une salade*	***raconter des salades*** vs. *#raconter une salade*	Figement morphologique
il a fait la soupe $\approx$ *la soupe a été faite par lui*	*il a fait la tête* vs. *#la tête a été faite par lui*	Figement syntaxique
les patates sont cuites $\approx$ *nous avons cuit les patates*	***les carottes sont cuites*** vs. *#on a cuit les carottes*	

'$\approx$': glissement de sens prévisible à partir du remplacement lexical

Expressions polylexicales – un vrai ***bric-à-brac*** 'amas d'objets hétéroclites'

- mots composés
  - ***cordon bleu*** 'excellent cuisinier', ***pomme de terre***
- termes complexes
  - ***mémoire vive*** 'RAM de l'ordinateur'
- entités nommées polylexicales
  - ***Mer Morte***
- constructions à verbe support
  - ***avoir besoin***, ***rendre visite***
- idiomes
  - ***ne pas être sorti de l'auberge*** 'ne pas avoir terminé avec des problèmes'
- proverbes
  - ***qui sème le vent récolte la tempête*** 'celui qui incite à la violence doit s'attendre à de fâcheuses conséquences'

Elles se rencontrent *tous les quatre matins* 'très souvent'

Fréquence des EP dans les textes

Jusqu'à 20% des mots d'un texte appartiennent à des expressions polylexicales.

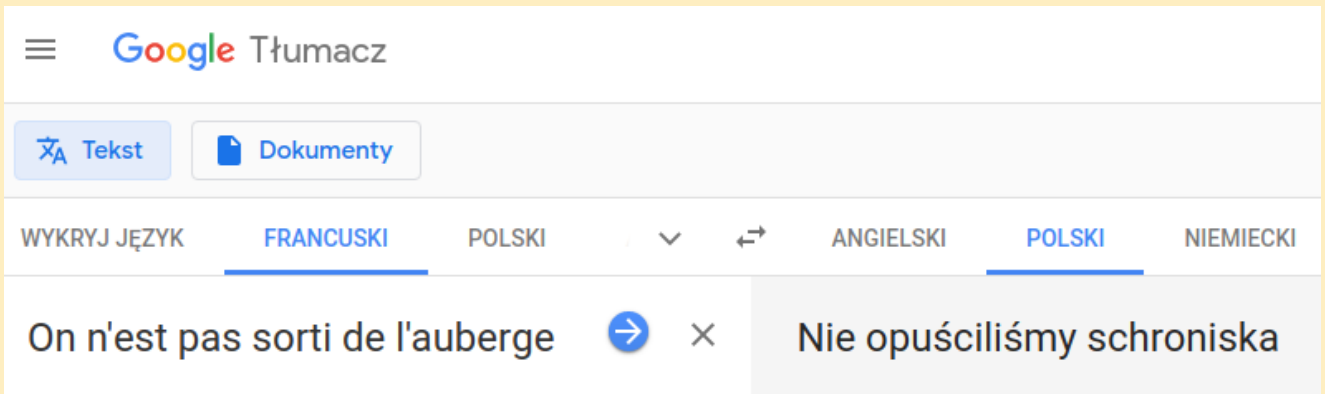
*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au **rédacteur-en-chef**, Monsieur **Jean-Marc Petit**.*

- Ici: 17 composants d'EP sur 30 mot du texte $\rightarrow$ 57%

Les *casse-pieds* 'qui ennuient/dérangent' de la linguistique computationnelle

Du fait de la non-compositionnalité, les EP constituent un défi pour des tâches du traitement sémantique.

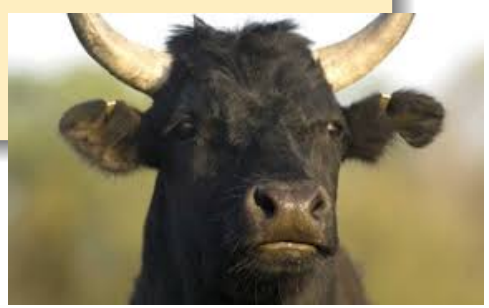
Traduction automatique



Traductions mot-à-mot ne captent pas le sens idiomatique.

Recherche d'information (RI)

- La tâche: pour une requête donnée (un ou plusieurs mots), trouver automatiquement tous les documents pertinents
- Moteurs de recherche: Google Search, Bing, Baidu, Yahoo!, ...
- Technique:
  - Tous les textes sont lemmatisés et **indexés**, i.e. représentés comme listes des mots qu'ils contiennent
  - Exemple: *les cornes et les pieds du taureaux* $\rightarrow$ {corne – 1, du – 1, et – 1, le – 2, pied – 1, taureau – 1}
  - Chaque mot d'une requête d'un utilisateur est recherchée dans l'indexe. Les textes contenant les bons mots sont pondérés et retournés en réponse.
- Problème avec les EP:
  - Un texte contient ***prendre le taureau par les cornes*** 'faire face aux difficultés plutôt que les fuir'
  - La requête d'un utilisateur contient *cornes d'un taureau*
  - Ce texte n'est pas pertinent, mais sera probablement proposé



Fouille d'opinion

- La tâche: prédire automatiquement la valence (positive, neutre ou négative) de l'opinion exprimée par un texte
- Exemples:
  - *Je soutiens les gilets jaunes. Je suis très respectueux de leur cause. Leur mouvement est compréhensible.*
  - *Il ne faut pas accepter ce geste. Rien ne le justifie ! Cela s'appelle une agression.*
  - *Tout cela va mal finir. La violence est partout dans les paroles et dans les actes.*
- Technique:
  - Les mots simples sont annotés avec une valence élémentaire: *respectueux* → 1, *agression* → -2, *justifier* → 1
  - Des règles locales modifient la valence élémentaire:
    - *très, grand* doublent la valence; *très respectueux* → 1*2 = 2; *grande agression* → -2*2= -4
    - La négation inverse la valence: *rien ne le justifie* → -1*1=-1

Fouille d'opinion – problèmes avec les EP

Texte	Valence calculée	Valence réelle
bras$_0$ *d'honneur*$_1$ 'geste impoli qui marque la reprobation'		
aimer$_2$ *à la folie*$_{-2}$ 'aimer énormément'		
boire$_0$ *comme un trou*$_0$ 'être alcoolique'		
avoir un coup$_0$ *de foudre*$_{-1}$ 'tomber amoureux d'un seul coup'		
avoir la pêche$_0$ 'être en forme'		
ne pas être dans son assiette$_0$ 'ne pas être en forme'		
la moutarde$_0$ *lui monte*$_0$ *au nez*$_0$ 'il se met en colère'		
avoir un poil$_0$ *dans la main*$_0$ 'être paresseux'		
tourner$_0$ *au vinaigre*$_0$'s'orienter vers la dispute'		
raconter des salades$_0$ 'dire n'importe quoi'		
en faire tout un fromage$_0$ 'exagérer'		
faire la tête$_0$ 'bouder'		
chercher la petite bête$_0$ 'regarder trop aux détails'		
ne pas être sorti$_0$ *de l'auberge*$_0$ 'continuer à avoir des difficultés'		
les carottes$_0$ *sont cuites*$_0$ 'la situation est compromise'		

Fouille d'opinion – problèmes avec les EP

Texte	Valence calculée	Valence réelle
bras$_0$ *d'honneur*$_1$ 'geste impoli qui marque la repropbation'	1	-2
aimer$_2$ *à la folie*$_{-2}$ 'aimer énormément'	0	4
boire$_0$ *comme un trou*$_0$ 'être alcoolique'	0	-2
avoir un coup$_0$ *de foudre*$_{-1}$ 'tomber amoureux d'un seul coup'	0	2
avoir la pêche$_0$ 'être en forme'	0	1
ne pas être dans son assiette$_0$ 'ne pas être en forme'	0	-1
la moutarde$_0$ *lui monte*$_0$ *au nez*$_0$ 'il se met en colère'	0	-1
avoir un poil$_0$ *dans la main*$_0$ 'être paresseux'	0	-1
tourner$_0$ *au vinaigre*$_0$'s'orienter vers la dispute'	0	-1
raconter des salades$_0$ 'dire n'importe quoi'	0	-1
en faire tout un fromage$_0$ 'exagérer'	0	-1
faire la tête$_0$ 'bouder'	0	-1
chercher la petite bête$_0$ 'regarder trop aux détails'	0	-1
ne pas être sorti$_0$ *de l'auberge*$_0$ 'continuer à avoir des difficultés'	0	-1
les carottes$_0$ *sont cuites*$_0$ 'la situation est compromise'	0	-1

Solutions

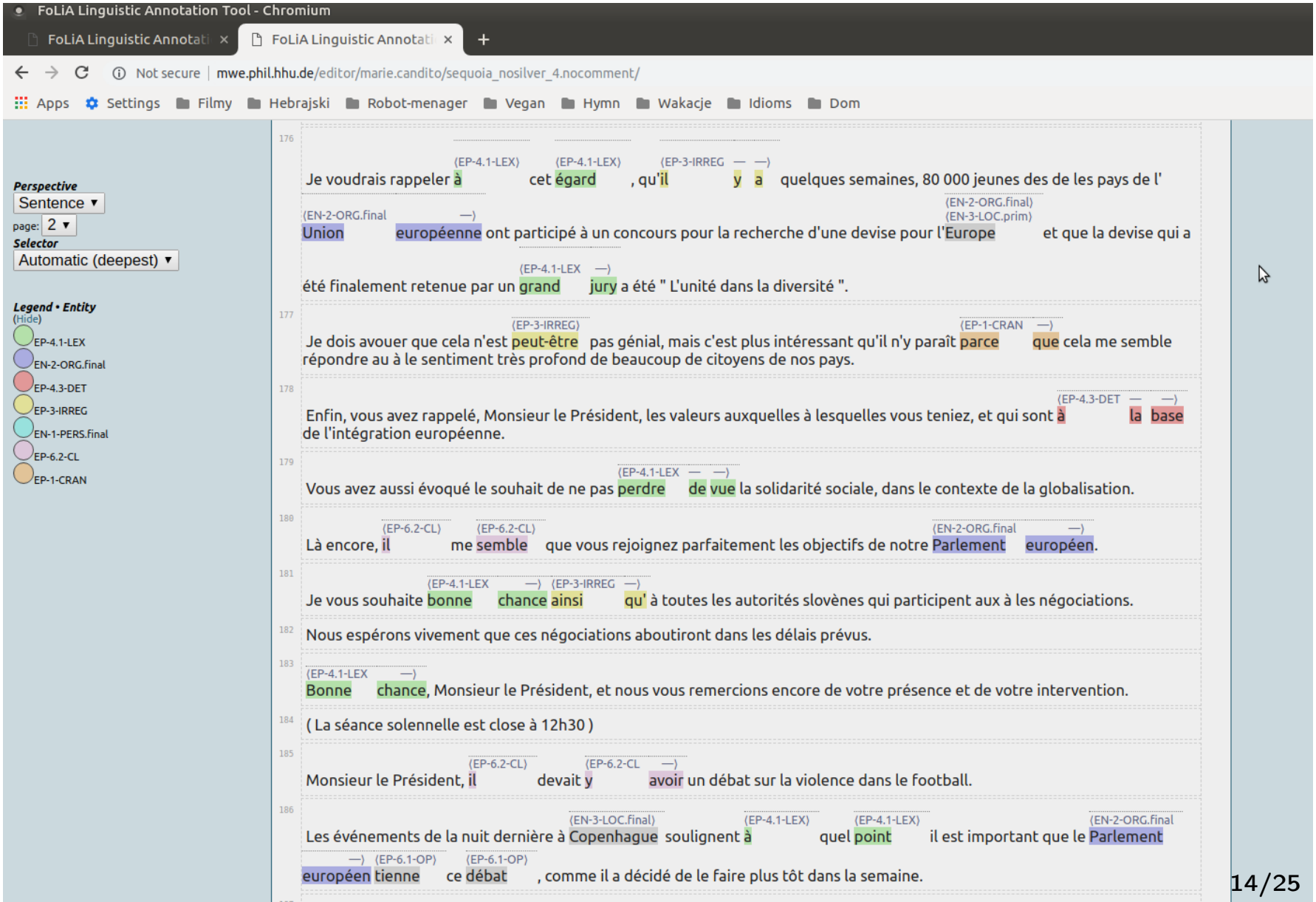
- Identifier automatiquement les EP dans le texte.
- Leur appliquer des traitements spéciaux:
  - traduction automatique
    - reformuler une EP avant la traduction
    - *il **boit comme un trou*** → *il est alcoolique* → *jest pijakiem*
  - recherche d'information
    - ne pas rajouter les composants d'une EP dans l'indexe du texte
    - rajouter l'expression en entier
    - *Le débat au parlement a **tourné au vinaigre*** → {au – 1, avoir – 1, débat – 1, le – 1, parlement – 2, tourner au vinaigre – 1}
  - fouille d'opinion
    - attribuer une valence à une EP en entier
    - *avoir un [**coup de foudre**]*$_2$

Rôle des linguistes

Créer des ressources linguistiques:

- lexiques
- corpus annotés

Annotation d'EP en corpus



Corpus PARSEME d'expressions polylexicales verbales (EPV)

Coopération internationale

- réseau scientifique européen PARSEME: 26 équipes nationales
- terminologie et méthodologie unifiées
- corpus de 26 langues, 8 millions de words, 113 mille EP annotées

Familles de langues

- **Balto-slaves**: bulgare (BG), croate (HR), lituanien (LT), polonais (PL), serbe (RS), slovène (SL), tchèque (CS)
- **Germaniques**: allemand (DE), anglais (EN), suédois (SV)
- **Romanes**: espagnol (ES), français (FR), italien (IT), portugais brésilien (PT), roumain (RO)
- **Autres**: arabe (AR), basque (EU), chinois (ZH), farsi (FA), grec (EL), hébreu (HE), hindi (HI), hongrois (HU), irlandais (GA), maltais (MT), turc (TR)

Typologie

Catégories universelles (valables dans toutes les langues)

- idiomes verbaux (verbal idioms : **VID**)

  ne pas être sorti de l'auberge

- constructions à verbe support (light-verb constructions : **LVCs**) - le nom exprime l'action/état, le verbe apporte seulement la morphologie (temps, mode, personne, etc.)

  rendre visite (**LVC.full**)
  entraîner *la* ***mort*** (**LVC.cause**)

Catégories quasi-universelles (valables dans plusieurs langues)

- verbes intrinsèquement pronominaux (inherently reflexive verbs : **IRVs**)

  s'apercevoir

- verbes à particule (verb-particle constructions : **VPCs**)

  (EN) *to* ***do in*** 'tuer' (VPC.full)
  (EN) *to* ***eat up*** 'manger tout' (VPC.semi)

- constructions multi-verbales (multi-verb constructions : **MVCs**) – en langues asiatiques

  (HI) ***kar le-na*** 'do take.INF'⇒'to do something (for one's own benefit)'

Guide d'anotation – *metez la main à la pâte* 'travaillez vous-mêmes'

Tu feras d'une pierre deux coups, si tu te comportes bien.

Tu feras d'une pierre deux coups

- Suivre l'arbre de décision
- S.1 [1HEAD] (YES): *fera* est la seule tête verbale
- S.2 [1DEP] (NO): *d'une pierre* et *deux coups* sont les deux dépendants lexicalisés
- VID.1 [CRAN] (NO): tous les composants existent comme mots indépendants
- VID.2 [LEX] (YES): *#tu feras de deux pierres quatre coup*
- Résultat: **VID**

Guide d'anotation

Tu feras d'une pierre deux coups, si tu te comportes bien.

si tu te comportes bien

- S.1 [1HEAD] (YES): *comportes* est la seule tête verbale
- S.2 [1DEP] (YES): *te* est le seul dépendant lexicalisé de *comportes*
- S.3 [LEX-SUBJ] (NO): le sujet *tu* n'est pas lexicalisé
- S.4 [CATEG] (reflexive clitic): *te* est un pronom reflexif
- IRV.1 [INHERENT] (NON): *comporte* peut apparaître tout seul
- IRV.2 [DIFF-SENSE] (YES): *se comporter* a un sens très différent que *comporter*
- Résultat: **IRV**

Tâche d'identification automatique d'EPV

C'est *tip-top* 'excellent' ou ça *ne casse pas trois pattes à un canard* 'c'est médiocre' ?

Evaluation d'une annotation automatique

- **Vrais** (V) – entités existantes dans le texte (annotées par un humain)
- **Positifs** (P) – entités identifiées par le système
- **Vrais positifs** (VP) – entités correctement identifiées par le système

Mesures d'évaluation

- **Précision**: $P = \frac{|VP|}{|P|}$
- **Rappel**: $R = \frac{|VP|}{|V|}$
- **F-mesure**: $F = \frac{2*P*R}{|P|+|R|}$

C'est **tip-top** ou ça **ne casse pas trois pattes à un canard** ?

Vrais (annotés par un humain)

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au **rédacteur-en-chef**, Monsieur **Jean-Marc Petit**.*

Positifs (identifiés par le système)

Si vous avez tant besoin de couper l'herbe sous le pied de quelqu'un, je vous proposerais de vous en prendre au rédacteur-en-chef, Monsieur Jean-Marc Petit.

- $|V| = ?$
- $|P| = ?$
- $|VP| = ?$
- $P = \frac{|VP|}{|P|} = ?$
- $R = \frac{|VP|}{|V|} = ?$
- $F = \frac{2*P*R}{|P|+|R|} = ?$

C'est **tip-top** ou ça **ne casse pas trois pattes à un canard** ?

Vrais

*Si vous **avez** tant **besoin** de **couper l'herbe sous le pied** de quelqu'un, je vous proposerais de **vous en prendre** au **rédacteur-en-chef**, Monsieur **Jean-Marc Petit**.*

Positifs

Si vous avez tant besoin de couper l'herbe sous le pied de quelqu'un, je vous proposerais de vous en prendre au rédacteur-en-chef, Monsieur Jean-Marc Petit.

- $|V| = 5$
- $|P| = 4$
- $|VP| = 2$
- $P = \frac{2}{4} = \mathbf{0{,}5}$
- $R = \frac{2}{5} = \mathbf{0{,}4}$
- $F = \frac{2*P*R}{|P|+|R|} = \frac{2*0{,}5*0{,}4}{0{,}5+0{,}4} = \frac{0{,}4}{0{,}9} = \mathbf{0{,}44}$

Campagne d'évaluation PARSEME (édition 1.2 en 2020)

- 7 équipes, 9 systèmes d'identification d'EPV, 14 langues
- Le corpus PARSEME pour chaque langue est divisé en 2 parties
  - (grand) corpus d'entraînement (TRAIN)
  - (petit) corpus d'évaluation en deux versions: annotée (TEST) et non-annotée (TEST.blind)
- Les systèmes automatiques s'entraînent sur le TRAIN
- Ils sont appliqués au TEST.blind
- Les organisateurs comparent leurs résultats au TEST et calculent P, R et F

On a fait du chemin 'on a progressé', mais *on n'est pas sorti de l'auberge* 'on n'en a pas fini avec des difficultés'

Meilleurs résultats pour toutes les langues

Système	P	R	F
MTLB-STRUCT (édition 1.2 2020)	0,71	0,69	**0,70**
Seen2Seen (édition 1.2 2020)	0,76	0,58	**0,66**
Meuilleur système dans l'édition 1.1 (2018)	0,66	0,51	**0,58**

Meilleurs résultats pour le français

Système	P	R	F1
TRAVIS-mono (édition 1.2 2020)	0,82	0,83	**0,82**
MTLB-STRUCT (édition 1.2 2020)	0,80	0,78	**0,79**
Meuilleur système dans l'édition 1.1 (2018)	0,72	0,53	**0,61**

Point barre! 'tout a été dit'

Conclusions

- Les EP sont fascinantes!
  - Elles font passer le message de manière succincte et efficace
  - Elle cachent des traces de l'histoire, des stéréotypes, des connotations surprenantes
  - Elles peuvent être très drôles
- Les EP sont difficiles
  - Elles sont difficiles à apprendre pour les locuteurs non natifs
  - Elles témoignent du degré la maîtrise d'une langue
  - Elles se comportent différemment que des combinaisons régulières de mots
  - Elles sont difficiles à identifier, analyser et traduire automatiquement
- Les EP très fréquentes
  - Elles couvrent jusqu'à 20% de mots d'un texte en langage naturel