WG3
Statistical Measures to characterise MWEs involving 'mordre' in French or 'bite' in English

Ismail El Maarouf[1] and Michael Oakes[2]

[1]Oxford University Press [2] University of Wolverhampton

Introduction

- **Context** Corpus Pattern Analysis (CPA; Patrick Hanks) and the Pattern Dictionary of English Verbs (http://pdev.org.uk)
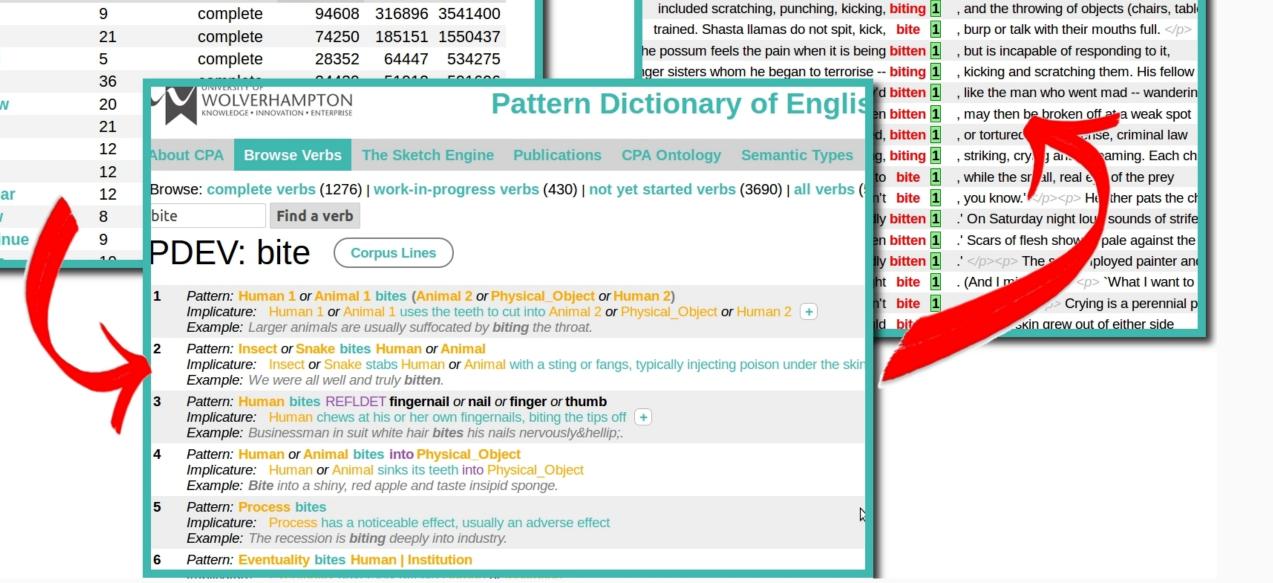- Approach statistical Measures for characterizing Multiword expressions. Focus on idiomatic patterns.
- Motivation Better insight of statistical features of idioms for automatic extraction
- Experiment Comparative study of statistical metrics on French *mordre* and English *bite*.
- PDEV: 1276 entries (1706) and 4601 (6755) patterns.

Pattern set	entries	patterns
Phrasal Verbs	195	506
Idioms	200	456
Phrasal and Idiom Overlap	17	28
Lexically Grounded	530	1268
Lexically Grounded (no-overlap)	460	844

pdev.org.uk: a bird's eye view

Statistical metrics for MWEs

Spread: Mean μ of all distances

$$\mu_{(X,Y)} = \frac{1}{n}\sum_{i=1}^{n} dist(X_i, Y_i)$$

Flexibility: Standard deviation σ of distances

$$\sigma_{(X,Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(dist(X_i, Y_i) - \mu_{(X,Y)})^2}$$

Diversity E as Entropy

$$E_{(X,Y)} = -\sum_{i=1}^{n} P_j log_2 P_j$$

Worked out example: 'bite', 'dog'

Text distance between collocations

Calculation

For $X = \{dog, dogs\}$, $Y = \{bite, bites, bit, bitten\}$, $i = \{1,2,3,4,5,6,7,8,9,10\}$

$j = \{$"that barks doesn't","that had been", "another. In", "to", "by a police","","his pet","are","always"$\}$

$$\mu_{X,Y} = \frac{(-4)+(-4)+3+(-2)+4+(-1)+3+(-2)+(-1)+(-2)}{10} = -0.6$$

$$\sigma_{X,Y} = \sqrt{\frac{(-4-(-0.6))^2 + (-4-(-0.6))^2 + (3-(-0.6))^2 + ...}{10}} = 2.76$$

$$E_{X,Y} = -((\frac{1}{10}\log_2\frac{1}{10}) + (\frac{1}{10}\log_2\frac{1}{10}) + ...) = 3.12$$

Results: English idioms

Collocate	Keyword	Freq	PMI	T-score	Log-Dice	Idiomaticity	Entropy	Mean span	Std. dev.
back	bite	87	5.914	10.380	5.549	0.989	0.338	1.057	0.277
bullet	bite	36	10.484	6.477	8.561	1	1.069	2.055	0.404
head	bite off	30	6.009	7.639	5.600	0.775	3.281	3.032	2.721
dust	bite	26	8.918	5.088	7.438	1	0.235	2.03	0.192
bug	bite	19	10.589	4.688	7.894	0.842	3.326	3.125	2.578

Results: French idioms

Collocate	Keyword	Freq	PMI	T-score	Log-Dice	Idiomaticity	Entropy	Mean span	Std. dev.
poisson	mordre	4	7.34	1.988	-2.222	0.75	0	1	0
hameçon	mordre	4	12.259	2	2.664	0.75	0	3	0
appât	mordre	2	10.475	1.413	894	1	0	3	0
vie	mordre	6	4.434	2.523	-5.126	1	1.792	2.833	0.372
doigt	mordre	21	9.387	4.789	-0.174	0.952	1.08	1.190	0.154
poussière	mordre	6	9.36	2.446	-0.204	1	0	1	0
queue	mordre	16	9.67	4.118	0.109	1	0.34	1.187	0.527
serpent	mordre	13	11.022	3.604	1.457	1	1.7	1.846	0.591

English patterns for 'bite' (BNC sample)

Pattern / Implicature	freq.	Perc.
Human 1 bites Human 2's head off	5	1.22
Human 1 speaks sharply and unkindly to Human 2		
Human bites REFLDET lip	8	1.96
Human grips his or her lip firmly with the teeth		
Human bites off more than Human can chew	4	0.98
Human undertakes a task that is too difficult for him or her to accomplish successfully		
Human bites the hand that feeds Human	5	1.22
Human attacks his or her benefactor		
Human or Institution bites the bullet	21	5.13
Human or Institution decides to do something necessary but unpleasant		
Human is bitten by the [MOD] bug	7	1.71
Human becomes very interested in [MOD]		
Human bites the dust	2	0.49
Human dies suddenly and violently		
Entity or Process bites the dust	8	1.96
Entity or Process comes to a sudden and unwelcome end		
Human bites [REFLDET] tongue	8	1.96
Human makes a desperate effort not to say what is on his or her mind		
Once bitten twice shy	3	0.73
An unpleasant experience causes someone to be more cautious in future		

French patterns for 'mordre' (FrTenTen sample)

Pattern and Implicature	freq.	perc.
Human ou le poisson mord (à l'hamecon ou à l'appat)	10	2
Human takes the bait (= is lured to do something that has bad consequences)		
Human mord la vie à pleines dents	6	1.2
Human enjoys life to the full [literally, bites life with full teeth]		
Human se mord les doigts	21	4.2
Human experiences a bitter time [literally, bites hisher fingers]		
Human 1 fait mordre la poussière à Human 2	6	1.2
Human 1 causes Human 2 to bite the dust (= to die) or to lose a challenge [the latter sense only in French]		
le serpent se mord la queue	16	3.2
Human is stuck in a State of affairs and cannot find a way out [literally, the snake bites his own tail]		
Human ne mord pas [NO OBJ]	6	1.2
Human does not bite (= is harmless)		

References and Acknowledgments

Hanks, Patrick. 2013.Lexical Analysis: Norms and Exploitations.Cambridge. MIT Press.
Oakes, Michael. 2012. Describing a Translational Corpus. In: M. P. Oakes and M. Ji (eds), Quantitative Methods in Corpus-Based Translation Studies: 115-148. Amsterdam: John Benjamins.

Conclusions and Perspectives

- Similar behaviour of metrics in French & English.
- Insights into literal translation VS idiomatic translation for Machine translation.
- Extend to more languages and more idioms
- Generate statistical profiles for MWEs automatically.