Measures of collocational strength and flexibility for the identification of MWEs

Patrick Hanks, Ismail El Maarouf, and Michael Oakes

patrick.w.hanks@gmail.com, i.el-maarouf@wlv.ac.uk, michael.oakes@wlv.ac.uk

Introduction

- Automatic identification of MWE
- Word association Measures (Pecina, 2008)
- Idioms in the British National Corpus

Research context

- Corpus Pattern Analysis (Hanks, 2013), DVC project.
- The Pattern Dictionary of English Verbs (http://pdev.org.uk)
- Representation and annotation of MWEs

Measuring the flexibility of MWE: definitions and worked-out example

Statistical formulas

Mean μ of all distances

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

Spread σ as the standard deviation 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

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

Text distance between collocations

`A *dog* that barks doesn't	bite	1	,' replied Antonio Navarro,
of the *dogs* that had been	bitten	2	and strayed: scared that th
in saliva when one animal	bites	3	another. In *dogs* , one of th
who had trained his *dog* to	bite	4	Arabs, and who informed u
> He was chased and	bitten	5	by a police *dog* and then a
t was saved when her *dog*	bit	6	him. The 22-year-
heltenham yesterday after	biting	7	his pet *dog* , which was atl
time by their own *dog* are	bitten	8	in the bedroom. In our bree
d. After that *dogs*	bit	9	me on the feet. Blood cam
/ herself that *dogs* always	bite	10	people, especially them. Tl

For $X = \{dog, dogs\}$, $Y = \{bite, bites, bit, bitten\}$, $i = \{1,2,3,4,5,6,7,8,9,10\}$ and $j = \{$*"that barks doesn't","that had been", "another. In", "to", "by a police","","his pet","are","always"*$\}$

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

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

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

Case study: MWEs of to bite

PDEV bite: 10 idioms

13 *Pattern:* Idiom. **Human 1 bites** Human 2's **head off**
Implicature: Human 1 speaks sharply and unkindly to Human 2
Example: *Just to **bite** their heads off.*

14 *Pattern:* Idiom. **Human bites lip**
Implicature: Human grips his or her lip firmly with the teeth +
Example: *He **bit** his lip but stood his ground.*

15 *Pattern:* Idiom. **Human bite off more than [[Human]] can chew**
Implicature: Human undertakes a task that is is too difficult for him or her to accomplish successfully
Example: *By aiming to depict Life in the 1990s, Kasdan has probably **bitten** off more than he can chew, but he*

16 *Pattern:* Idiom. **Human bites** the **hand that feeds [[Human]]**
Implicature: Human attacks his or her benefactor +
Example: *It is hard to **bite** the hand that feeds you.*

17 *Pattern:* Idiom. **Human | Institution bites the bullet**
Implicature: Human *or* Institution decides to do something necessary but unpleasant +
Example: *So, this week, Priddle **bit** the bullet.*

18 *Pattern:* Idiom. **Human is bitten by the [MOD] bug**
Implicature: Human becomes very interested in [MOD]
Example: *Chubby, bubbly jazzman Fats Waller was among the first to really get **bitten** by the London bug.*

19 *Pattern:* Idiom. **Human bites the dust**
Implicature: Informal. Human dies suddenly and violently
Example: *They **bite** the dust with lead in their bellies.*

20 *Pattern:* Idiom. **Entity** *or* **Process bites the dust**
Implicature: Informal. Entity *or* Process comes to a sudden and unwelcome end
Example: *If so, then we must freely admit that another time-honoured tradition of British self-restraint has very r*

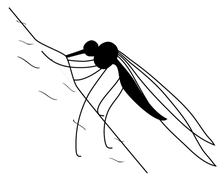
21 *Pattern:* Idiom. **Human bites** REFLDET **tongue**
Implicature: Human makes a desperate effort not to say what is in his or her mind
Example: *It's all very well telling someone to **bite** their tongue and not fight back.*

22 *Pattern:* Idiom. **once bitten twice shy**
Implicature: an unpleasant experience causes someone to be more cautious in future
Example: *This time around it is a case of `once **bitten**, twice shy' and their doubt is not simple but compound.*

Analysis of the flexibility of idioms in BNC50

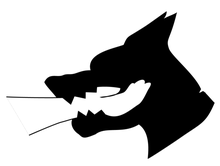
X,Y	Freq.	μ	σ	E
bite, bullet	9	3	0	0
bite, back	3	1	0	0
bite, feed*	5	4	0	0
bite, off*	6	4	0	1.057
bitten, bug	4	3.75	1.5	2

* including variants



Examples found for {bite, bug}

Waller was among the first to really get	bitten	18	by the London bug. In 1939 he fashi
t their badly burned faces. He was	bitten	18	by the flying bug at the age of four w
and in Africa, when he was, like Vine, `	bitten	18	by the bug of the ocean floor'. In 197
with his prize heifers. But he had been	bitten	18	by the newspaper bug, and various g



Examples found for {bite, feeds}

uneasily with service delivery. It is hard to	bite	16	the hand that feeds you. There is a da
ctual home. They are not normally going to	bite	16	the hand that feeds them. The questio
leceptive because they may then suddenly	bite	16	the hand that holds them. This behavi
principal apologist for Official Nationality	biting	16	the hand that fed him must have convi
subservience, the insider can find it hard to	bite	16	the hand that feeds and reveal any unh

Perspectives

- Combine measures with structural morpho-syntactic information clues and with word association measures such as MI.
- Study sensitivity of Entropy with other units e.g. characters, words.
- Use a Machine Learning classifier for the discrimination of MWEs.
- Experiment with other languages (less fixed word order)

Acknowledgements

Image credits to Graphist Illustrator Marion Dugalès, https://www.linkedin.com/profile/view?id=282216480
This work is partly supported by AHRC DVC-AH J005940/1, 2012-2015

References

- Pattern Dictionary of English Verbs **http://pdev.org.uk**

- Kenneth W. Church and Patrick Hanks. 1989. *Word Association Norms, Mutual Information and Lexicography.* Proc. ACL: 76-83.
- Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* MIT Press.
- Adam Kilgarriff, Pavel Rychly, Pavel Smrz and David Tugwell. 2004. The Sketch Engine. Proc. Euralex: 105-116.
- Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction.* PhD thesis, Charles University in Prague.
- Michael P. Oakes. 2012. *Describing a Translational Corpus.* In: Oakes, M. P. and Ji, M., *Quantitative Methods in Corpus-Based Translation Studies.* John Benjamins: 115-148.
- Frank Smadja. 1993. *Retrieving collocations from text: Xtract.* Computational Linguistics 19: 143-177.