How (not) to analyse multiword expressions

JAMIE Y. FINDLAY
jamie.findlay@ling-phil.ox.ac.uk

University of Oxford

Linguistics Oberseminar

25 June, 2018

Outline

UNIVERSITY OF OXFORD
FACULTY OF LINGUISTICS, PHILOLOGY AND PHONETICS

What are multiword expressions?

Explananda

Three kinds of theory

One size fits all?

What are multiword expressions?

Multiword expressions

- Necessary and sufficient conditions hard to give.
- Cover term used to describe a wide-ranging, heterogeneous group: idioms, phrasal verbs, light verb constructions, compounds, some proper names, . . .
- *kick the bucket, pull strings, rely on, consist of, take a break, have a shower, washing machine, cat food, New York, Jack the Ripper, . . .*
- Tension between word-like and phrase-like properties.

Multiword expressions

- Some common characteristics:
  - **not (straightforwarldy/entirely) compositional**
    - their meaning is more than or different from the sum of their parts
  - **parts must appear together**

    (1) a. #Those are some impressive strings!
    b. #We mustn't pull Chris's good nature.

- So are they just big words?

Words with spaces

- The *monolexemic approach*: MWEs are single morphological objects, i.e. "words with spaces" (Sag et al. 2002)
- [kick the bucket]$_{\mathrm{V}}$

Flexibility

- Major problem: many MWEs are far more flexible than words.

Flexibility

- **Morphological flexibility (inflection)**:

(2) a. When Sandy [kicks the bucket], . . .
b. When Sandy and Kim [kick the bucket], . . .
c. Since they've [kicked the bucket], . . .

Flexibility

- **Syntactic flexibility**:

(3) a. I [pulled strings] to get you into the club.
b. [Strings were pulled] to get you into the club.
c. The [strings that I pulled] got you into the club.

Flexibility

- **Modification**:

(4) a. Alex kicked the proverbial bucket.
b. We leave no digital stone unturned, we poke and prod every nook and cranny of the Interwebz.
c. Emotiva has lots of bigger fish to fry at the current time.

Multiword expressions

- Tension between the two sides of MWEs:
  - **Word-like/Unitary**:
    Non-compositional meaning
    Parts must appear together
  - **Phrase-like/Divided**:
    Flexibility (parts can be altered, separated, re-ordered)
- Monolexemic approach comes down entirely on one side: focuses on unitary nature of MWEs while neglecting their phrase-like properties.

Explananda

Explananda

- Idiomaticity
- The real words problem
- Flexibility
- Modification
- Psycholinguistic findings

Idiomaticity

- Three kinds of idiomaticity/idiosyncratic behaviour (Baldwin & Kim 2010):
  - Semantic
  - Syntactic
  - Lexical

Semantic idiomaticity

- Non-compositionality taken as definitional by many.
- E.g. *painting the town red* has nothing to do with redecorating municipal buildings.
- Any analysis of MWEs must be able to account for how MWEs come to have the idiosyncratic semantics they do.

Syntactic idiomaticity

- Some MWEs ambiguous between a literal and idiomatic reading:

(5) *Kira pulled a rabbit out of the hat*.

= [literal] Kira extracted an actual rabbit from a real hat (perhaps she is a magician)

= [idiomatic] Kira did something unexpected but ingenious to solve a problem.

Syntactic idiomaticity

- But some don't even have a literal interpretation, because their syntax is archaic or otherwise weird:

(6) a. We [$_{\text{VP}}$ tripped [$_{\text{NP}}$ the [$_{??}$ light fantastic]]] all night long.

b. This was [$_{??}$ by and large] a success.

c. Cars, trucks and buildings located 300 meters away were all blown to [$_{??}$ kingdom come].

d. There were people running [$_{??}$ every which way].

Lexical idiomaticity

- Other MWEs don't receive a literal parse because they contain words which don't/no longer appear outside of the MWE ('cranberry words'):

(7) a. It's good to be on [terra firma] again!
b. Two jet-skiers who [ran amok] in Portsmouth Harbour have been fined for their actions.
c. They [took umbrage] at how much time Gregory spent hobnobbing at his vacation home on Nantucket.

The real words problem

- Although they can contain words which do not exist outside of the expression, MWEs nonetheless overwhelmingly contain words that do.

- MWEs made up entirely of ‘cranberry words’ are conceivable, but not attested:

(8) *flargbliff* > *flarg bliff*

a. Edward flargs bliff with some aplomb.
b. You shouldn’t flarg bliff in front of the boss.
c. The bliff she flarged was most impressive.

The real words problem

- What is more, when parts of an MWE have an irregular paradigm in their literal usage, they will generally retain this irregular paradigm within the MWE as well:

  (9) a. Quark comes/came a cropper.
  b. Miles loses/lost his cool.

- It would be desirable if it were not left as a coincidence that MWEs are at least partly made up of pre-existing words from the language.

Flexibility

- Just as there are various kinds of idiomaticity, so too there are several ways in which MWEs can be ‘flexible’.
- We have already seen morphological and syntactic flexibility.
- (There are varying degrees of each.)

Lexical flexibility

- Some also exhibit a looser degree of fixedness when it comes to the words they contain:

  (10) To give someone a kick up the arse/backside/ butt/derrière/. . .

Modification

- Three kinds of modification (Ernst 1981):
  1. Internal
  2. External
  3. ‘Conjunction modification’

Internal modification

- Parts of idioms can have idiomatic meanings of their own; we call these idioms *decomposable* (Nunberg et al. 1994 call them *idiomatically combining expressions*).

- These parts can then be modified directly:

  (11) Tom won't pull family strings to get himself out of debt.

  (12) Maybe by writing this book I'll offend a few people or touch a few nerves.
  (= I will upset a few people or annoy someone in a few ways.)
  (≠ I will cause the same irritation multiple times.)

External modification

- Modifiers are still possible with non-decomposable idioms, but are often interpreted as scoping more widely than their position would suggest:

  (13) The President doesn't have an economic leg to stand on.
  (= Economically, the President doesn't have a leg to stand on.)

  (14) Britney Spears came apart at the mental seams.
  (= Mentally, Britney Spears came apart at the seams.)

- Cf. '*occasional*-type' adjectives (Bolinger 1967; Gehrke to appear):

  (15) An occasional sailor comes into the bar.
  (= Occasionally, a sailor comes into the bar.)

Conjunction modification

- The modifier applies to the *literal* meaning of the idiom part (which can then be interpreted as expressing an additional proposition, either about the literal or idiomatic meaning):

  (16) Shepard enjoys pulling Jack's tattooed leg.

  (17) With the recession, oil companies are having to tighten their Gucci belts.

‘Extended’ uses

- Conjunction modification is related to a broader family of extended uses of idioms:

  (18) If you let this cat out of the bag, a lot of people are going to get scratched.

  (19) A: Can I bounce an idea off you?
  B: All right, but don’t throw it too hard, I can hardly think straight as it is!

  (20) Alastair tried to pull some strings for me, but they snapped.

- These seem to necessitate ‘reactivating’ the metaphor behind the idiom.

Pronominalisation

- In some contexts, part of an idiom can be replaced with a proform:

  (21) (*)Although the F.B.I. kept tabs on Jane Fonda, the C.I.A. kept them on Vanessa Redgrave. (Bresnan 1982: 49)

  (22) Pat tried to break the ice, but it was Chris who succeeded in breaking it. (Nunberg et al. 1994: 502)

- There is a question mark over how acceptable some of these examples are, and, if acceptable, whether they are instances of 'extended' uses.

Psycholinguistic findings

- Swinney & Cutler (1979): there is no special ‘idiom mode’ of comprehension which our minds switch into when confronted with idiomatic material.
- At the same time, idiomatic meanings are processed faster and in preference to literal ones (Estill & Kemper 1982; Gibbs 1986; Cronk 1992; i.a.).
- This might suggest there is a difference in their representation.

Three kinds of theory

Definitions of ‘word’

- Di Sciullo & Williams (1987): there are (at least) three ways of thinking about what a word is:
  1. Morphological object
  2. Syntactic atom
  3. Listeme
- At which level(s) do we represent the multiplicity of MWEs?

The space of possible theories

- Eight possibiltiies:

	A	B	C	D	E	F	G	H
Morphological Object	–	–	–	–	+	+	+	+
Syntactic Atom	–	+	–	+	–	+	–	+
Listeme	–	–	+	+	–	–	+	+

The space of possible theories

- Remove monolexemic theories:

	A	B	C	D	E	F	G	H
Morphological Object	–	–	–	–	+	+	+	+
Syntactic Atom	–	+	–	+	–	+	–	+
Listeme	–	–	+	+	–	–	+	+

The space of possible theories

- [Listeme +] $\rightarrow$ [Syntactic Atom +], so G is logically incoherent:

	A	B	C	D	E	F	G	H
Morphological Object	–	–	–	–	+	+	+	+
Syntactic Atom	–	+	–	+	–	+	–	+
Listeme	–	–	+	+	–	–	+	+

The space of possible theories

- [+, +, +] = H = lexical ambiguity/compositional approach (e.g. Kay et al. 2015; Lichte & Kallmeyer 2016; Bargmann & Sailer 2018)
- [+, −, −] = E = construction-based approach (e.g. Abeillé 1995; Jackendoff 1997; Findlay 2017)
- [+, +, −] = F = semantic/two-step approach (e.g. Pulman 1993; Egan 2008; Kobele 2012)

Lexical ambiguity

- The idea: treat MWEs just like other complex expressions, their meanings computed from the meanings of their parts and their syntax. So we need special meanings for the parts.
- Motivated by the decomposability facts:
  if *spill the beans* $\approx$ *divulge the secrets*,
  then *spill* $\approx$ *divulge* and *beans* $\approx$ *secrets*
- MWEs aren't really very special at all!

Lexical ambiguity

- Two versions: homonymy vs. polysemy.
- Are there two *spill*s, or just one with multiple meanings?

Lexical ambiguity: strengths

- A number of immediate advantages of treating the words in MWEs as normal words:
  - Words are the natural locus of idiosyncratic meaning, so semantic and lexical idiomaticity are no problem.
  - Morphological and syntactic flexibility fall out too: whatever your theory of morphology or syntax is, it applies just as usual.
  - Internal modification, too, is unsurprising: one can apply modifiers to the idiomatic meanings directly.
  - Because the parts are separate, nothing stops external modifiers appearing inside a MWE.

Lexical ambiguity: strengths

- These theories are at least in principle capable of handling the pronominalisation data, too: since each part of the MWE contains its own idiomatic meaning, it is not incoherent for other parts to be omitted.
- But not altogether clear how to toe the line between over- and under-generation. See below on the ‘collocational challenge’.

Lexical ambiguity: no man's land

- Syntactic idiomaticity is surprising on this account: we expect idiom words to behave just like ordinary words.
  - Pretty easy to fix with bespoke phrase-structure rules (or the equivalent).
  - But now MWEs have leaked out of the lexicon into the syntax too . . .

Lexical ambiguity: problems

- Conjunction modification and any extended use is impossible to make sense of: idiom words are ambiguous; they can't have both meanings simultaneously.
  - At least in the polysemy-based version of the theory, the literal meaning is still 'around' in some sense . . .
- Nothing to say about the psycholinguistic findings: if MWEs are formally identical to literal, compositional expressions, then why are they processed faster?

Lexical ambiguity: problems

- The homophony-based approach has no answer to the real words problem.
- At least the polysemy-based approach explains why we get the same inflectional paradigm (*come*∼*came a cropper*, etc.).
- However, ‘cranberry words’ pose a problem: if *cropper* has no synchronic existence outside of the MWE, then clearly idiom words need not be polysemes of existing words. And if arbitrary non-words are allowed in MWEs, then it is hard to see how they can be ruled out in general.

Lexical ambiguity: the collocational challenge

- One general problem for this approach: how to constrain idiom words so that they don't appear alone?

(23) a. #Those are some impressive strings!
(≠ . . . some impressive connections.)
b. #We mustn't pull Chris's good nature.
(≠ . . . exploit Chris's good nature.)

- Bargmann & Sailer (2018) refer to this as the *collocational challenge*.

Lexical ambiguity: the collocational challenge

- This is usually achieved via some kind of mutual selectional restriction, so that *pull* needs to be in the correct relation to *strings*, and vice versa.
- The challenge is in finding an appropriate level at which to describe this relation.

Lexical ambiguity: the collocational challenge

- It can't be in terms of grammatical functions, since there are idioms which passivise:

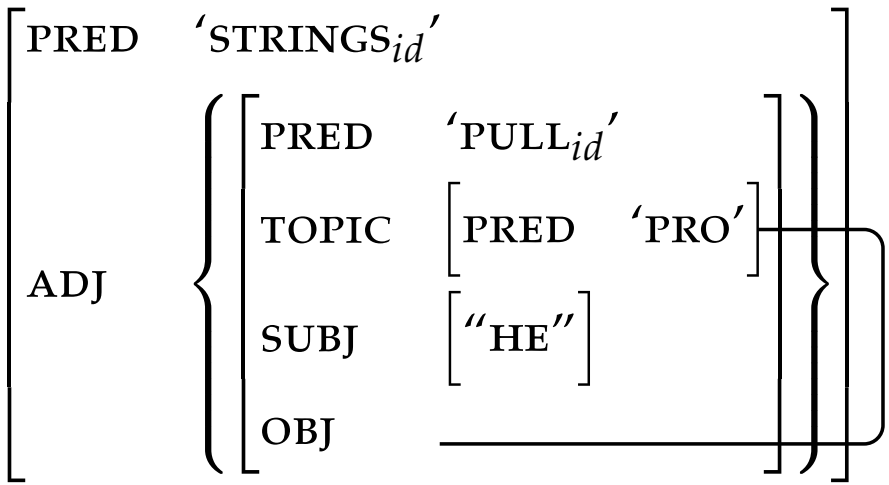
  (24) Strings were pulled for you, my dear. Did you really think the Philharmonic would take on a beginner like you?

- So maybe we should impose some restriction at the level of argument structure instead.

Lexical ambiguity: the collocational challenge

- But relative clauses pose a problem:

  the strings he pulled (for me)



- It is generally assumed that there is no direct syntactic or semantic relationship between the head noun of a relative clause and the predicate inside it.

Lexical ambiguity: the collocational challenge

- Instead, a relative pronoun, overt or covert, is taken to be the argument of the within-clause predicate, and this is related anaphorically to the head noun.

- So the two parts are only related by coreference. But this is too loose to serve as a general characterisation:

  (25) #Those are some impressive strings$_i$ – you should pull them$_i$ for me!

Lexical ambiguity: the collocational challenge

- We can give the following disjunctive description of the relation between idiomatic *pull* and *strings*:

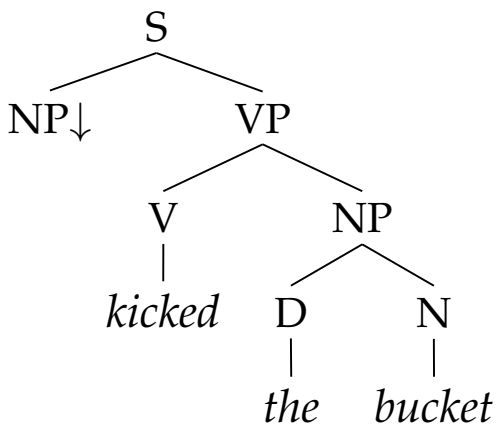
  (26) Either $strings_{id}$ is the internal argument of $pull_{id}$, or it is modified by an adjunct headed by $pull_{id}$ which has a pro-form coreferential with $strings_{id}$ as its internal argument.

- This is accurate, but smacks of a missed generalisation.

Construction-based approach

- The construction-based approach claims that MWEs are stored *en bloc*, but have internal structure.
- Essentially precompiled bits of syntax (associated with semantics):

(27)



Construction-based approach: strengths

- Once again, MWEs are directly encoded in the grammar, so semantic and lexical idiomaticity are straightforwardly handled.
- Since lexical entries are explicitly syntactic, syntactic idiomaticity is also no problem.
- The existence of sub-parts means they can be morphologically distinct, and also associated with separate meanings, allowing for internal modification.
- This kind of theory needs to be coupled with a formalism that allows the structure to be manipulated (e.g. by movement or adjunction) in order to allow for syntactic flexibility (and the various kinds of modification).

Construction-based approach: strengths

- The psycholinguistic findings are better explained in this kind of theory.
- A literal parse of *kick the bucket* requires looking up three items in the lexicon: *kick*, *the*, and *bucket*. The idiomatic parse only involves one: *kick the bucket*.

Construction-based approach: problems

- Like the lexical ambiguity approach, the construction-based approach falters when it comes to extended uses of idioms.
- It also has no reply to the real words problem, for much the same reasons as above.
- The pronominalisation facts are also harder to explain without resorting to mass ambiguity (e.g. one entry for *break the ice* and another for *break it*).

Semantic approach

- Another approach is to ignore the syntactic representation entirely, and focus solely on the meaning.
- Composition proceeds as usual, but then the output is inspected to see whether it matches an entry in an idiom list somewhere.
- Either we inspect a semantic representation and map it to a new one (Pulman 1993; Kobele 2012), or we interpret a meaning through some ‘pretence’ – a mapping between situations rather than representations (Egan 2008).
- Idioms are essentially conventions about meaning: ‘when we talk about kicking buckets, we are actually talking about dying’.

Semantic approach: strengths

- Like both other approaches, the semantic approach can easily handle semantic idiomaticity – that is precisely what it is designed to do.
- It also shares with the lexical ambiguity approach the fact that a lot of things come for free simply by assuming MWEs are constructed in the usual way, using the normal rules of the grammar.

Semantic approach: strengths

- Uniquely, these kinds of approaches can handle the extended uses, and highly distorted versions of expressions: as long as the right meaning gets across, it doesn't matter which words are used.

(28) a. Awww, I thought we'd snag at least one before the feline escaped from the bag.
b. Good gawd its another porcine flyer.
c. Then the manure really entered the ventilation system.

(29) Alastair tried to pull some strings for me, but they snapped.

Semantic approach: problems

- These kinds of approaches need a literal parse before they can get started, so lexical or syntactic idiomaticity is beyond their scope.
- Also, because they make no reference to syntax, such approaches have no way to constrain the syntactic flexibility of MWEs.
- Runs against psycholinguistic findings: processing of idioms should be more costly and slower if it involves a two-step process, and the literal parse always precedes the idiomatic one. But this isn't the case.

One size fits all?

Different theories for different MWEs?

- Different kinds of MWE, and different uses of the same expression, behave in different ways.
- Perhaps we want different theories for each of them.

Different theories for different MWEs?

- **Monolexemic**: fixed phrases like *by and large* or *bon appétit* – often very inflexible, and the source of syntactic idiomaticity.
- **Construction-based**: non-decomposable idioms, which require some internal structure to account for inflection and the presence of modifiers, but have limited syntactic flexibility and only a holistic meaning.
- **Lexical ambiguity**: decomposable idioms, where the meaning is distributed across the parts, and syntactic flexibility tends to be high.
- **Semantic**: for some highly lexically flexible idioms, as well as for extended uses where the metaphorical mapping is reactivated.

Different theories for different MWEs?

- But even the highly fixed phrases can be ‘brought to life’:

(30) a. A: By and large, the economy seems to be doing well.
B: By but not so large: have you seen the latest unemployment figures?
b. But by and indeed large, the Space-Sim has long since perished.

Different theories for different MWEs?

- Even the foreign borrowings:

(31) a. Bon bloody appetit!
[literal: said of some Halloween snacks decorated with fake blood]

b. Bon appétit, Marik! Bon bloody appétit!
[expressive: said in anger/frustration]

c. Mine, it appeared, had suffered from the full thud of my five-eleven frame hitting terra extremely firma.

Different theories for different MWEs?

- These expressions are not obviously decomposable (although the foreign borrowings are with some knowledge of the source language), so not good candidates for the lexical ambiguity approach?
- No particular motivation for lexical ambiguity approach for non-decomposable idioms either.
- The only independent motivation seems to be the decomposability facts, but the construction-based approach can handle that just as well.

Different theories for different MWEs?

- The extended uses can only be explained by a semantic approach, but we need one of the two other theories to handle the ‘core’ uses.
- The construction-based approach is capable of serving this role, and seems to me to fit our intuitions better.
- Just like the monolexemic approach, the lexical ambiguity theories come down entirely on one side of the tension between the unitary and divided natures of MWEs – this time emphasising their phrase-like properties above all else.
- Under this approach, MWEs have no unity; they are merely conspiracies of multiple, separate lexical items.

Conclusion

- MWEs exhibit a tension between word-like and phrase-like properties – this should be represented in any theory which purports to explain them.
- The construction-based approach achieves this better than the lexical ambiguity approach.
- But we still need the semantic approach to handle the extended, playful uses.

References I

Abeillé, Anne. 1995. The flexibility of French idioms: a representation with Lexicalized Tree Adjoining Grammar. In Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schreuder (eds.), *Idioms: structural and psychological perspectives*, Hove, UK: Lawrence Erlbaum.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing (2nd edn.)*, 267–292. Boca Raton, FL: CRC Press.

Bargmann, Sascha & Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: insights from a multi-lingual perspective*, 1–29. Berlin, DE: Language Science Press.

Bolinger, Dwight. 1967. Adjectives in English: attribution and predication. *Lingua* 18. 1–34.

Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan (ed.), *The mental representation of grammatical relations*, 3–86. Cambridge, MA: MIT Press.

Cronk, Brian C. 1992. The comprehension of idioms: The effects of familiarity, literalness, and usage. *Applied Psycholinguistics* 13. 131–146.

Di Sciullo, Anna Maria & Edwin Williams. 1987. *On the definition of word* (Linguistic Inquiry monographs 14). Cambridge, MA: MIT Press.

Egan, Andy. 2008. Pretense for the complete idiom. *Noûs* 42(3). 381–409.

Ernst, Thomas. 1981. Grist for the linguistic mill: idioms and 'extra' adjectives. *Journal of Linguistic Research* 1(3). 51–68.

Estill, Robert B. & Susan Kemper. 1982. Interpreting idioms. *Journal of Psycholinguistic Research* 11(6). 559–568.

Findlay, Jamie Y. 2017. Multiword expressions and lexicalism. In Miriam Butt & Tracy Holloway King (eds.), *Proceedings of the LFG17 Conference*, 209–229. Stanford, CA: CSLI Publications. `http://web.stanford.edu/group/cslipublications/cslipublications/LFG/LFG-2017/lfg2017-findlay.pdf`.

References II

Gehrke, Berit. to appear. Multiple event readings and *occasional*-type adjectives. In Patricia Cabredo Hofherr & Jenny Doetjes (eds.), *The Oxford handbook of grammatical number*, Oxford, UK: Oxford University Press.

Gibbs, Raymond W., Jr. 1986. Skating on thin ice: Literal meaning and understanding idioms in context. *Discourse Processes* 9. 17–30.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Kay, Paul, Ivan A. Sag & Daniel P. Flickinger. 2015. A lexical theory of phrasal idioms. Unpublished ms., CSLI, Stanford. `http://www1.icsi.berkeley.edu/~kay/idiom-pdflatex.11-13-15.pdf`.

Kobele, Gregory M. 2012. Idioms and extended transducers. In Chung-hye Han & Giorgio Satta (eds.), *Proceedings of the 11th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+11)*, 153–161. `http://www.aclweb.org/anthology/W12-4618`.

Lichte, Timm & Laura Kallmeyer. 2016. Same syntax, different semantics: a compositional approach to idiomaticity in multi-word expressions. In Christopher Piñón (ed.), *Empirical issues in syntax and semantics 11*, 111–140. Paris, FR: Colloque de Syntaxe et Sémantique à Paris (CSSP). `http://www.cssp.cnrs.fr/eiss11/eiss11_lichte-and-kallmeyer.pdf`.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Pulman, Stephen G. 1993. The recognition and interpretation of idioms. In Cristina Cacciari & Patrizia Tabossi (eds.), *Idioms: processing, structure, and interpretation*, 249–270. London, UK: Lawrence Erlbaum.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword Expressions: a pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Compuational Linguistics (CICLing-2002)*, 1–15. Mexico City, MX.

Swinney, David A. & Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior* 18(5). 523–534.