Computational considerations of comparisons and similes

Vlad Niculae
University of Wolverhampton
vlad@vene.ro

Victoria Yaneva
University of Wolverhampton
v.yaneva@wlv.ac.uk

Abstract

This paper presents work in progress towards automatic recognition and classification of comparisons and similes.

Among possible applications, we discuss the place of this task in text simplification for readers with Autism Spectrum Disorders (ASD), who are known to have deficits in comprehending figurative language.

We propose an approach to comparison recognition through the use of syntactic patterns. Keeping in mind the requirements of autistic readers, we discuss the properties relevant for distinguishing semantic criteria like figurativeness and abstractness.

1 Introduction

Comparisons are phrases that express the likeness of two entities. They rely on specific patterns that make them recognisable. The most obvious pattern, __ be *like* __, is illustrated by the following example, but many subtler ways of building comparisons exist:

> "He was like his father, except he had a crooked nose and his ears were a little lopsided." (In "Black cat" by Alex Krill)

Similes are a subset of comparisons. The simile is a figure of speech that builds on a comparison in order to exploit certain attributes of an entity in a striking manner. According to the Oxford English Dictionary, what sets a simile apart from a comparison is that it compares "one thing with another thing of a different kind"[1].

[1]"simile, n. a figure of speech involving the comparison of one thing with another thing of a different kind, used to make a description more emphatic or vivid (e.g. as brave as a lion)" OED Online. June 2004. Oxford University Press. 06 February 2013 http://dictionary.oed.com/.

A popular example by Charles Dickens is:

> "Mrs. Cratchit entered: flushed, but smiling proudly: with the pudding, like a speckled cannon-ball, so hard and firm, (...)" (In "A Christmas Carol" by Charles Dickens)

The comparison between a Christmas pudding and a cannon-ball is so unexpected, as delicious deserts are not conventionally associated with cannon-balls (or any kind of metal objects), that the author needs to clarify the resemblance by adding "so hard and firm" right after the simile. Intuitively, the OED definition is confirmed by these two examples: *a Christmas pudding* and *a cannon-ball* are things of different kinds, whereas *he* and *his father* are things of the same kind (namely, human males). As we shall see, the borderline which divides some similes and fixed expressions is the degree of conventionality. Many other phrases used by Dickens in "A Christmas Carol" also link two notions of different kinds: Old Marley was "as dead as a doornail" and Scrooge was "as hard as flint" and "as solitary as an oyster". In these cases, however, the link between the two entities is a pattern repeated so many times that it has consequently lost its innovativeness and turned into a dead metaphor ("as dead as a doornail") or a conventional simile (sections 4.1, 5.4.2).

The scholarly discussion of the simile has been controversial, especially with respect to its relative, the metaphor. The two were regarded as very close by Aristotle's *Rhetoric*: "The simile, also, is a metaphor, the difference is but slight" (Aristoteles and Cooper, 1932). However, modern research has largely focused on metaphor, while the simile suffered a *defiguration*, described and argued against by Bethlehem (1996): in order to support the idea that the metaphor embodies the essence of figurativeness, the simile was gradually stripped of

its status as figure of speech.

Metaphor is defined as "a word or phrase applied to an object or action to which it is not literally applicable"[2].

In other words, a metaphor links features of objects or events from two different, often incompatible domains, thus being a "realization of a cross-domain conceptual mapping" (Deignan, 2005). We are interested in the parallel between similes and metaphors insofar as it points to an overlap. There are types of similes that can be transformed into equivalent metaphors, and certain metaphors can be rewritten as similes, but neither set is included in the other. This view is supported by corpus evidence (Hanks, 2012) and contradicts reductionist *defiguration* point of view, in a way that Israel et al. (2004) suggest: some metaphors express things that cannot be expressed by similes, and vice versa.

In computational linguistics, similes have been neglected in favour of metaphor even more than in linguistics[3] , despite the fact that comparisons have a structure that makes them rather amenable to automated processing. In sections 2 we discuss one motivation for studying comparisons and similes: their simplification to language better suited for people with ASD. Section 3 reviews related work on figurative language in NLP. In section 4 we present the structure of comparisons and some associated patterns, emphasising the difficulties posed by the flexibility of language. Section 5 describes computational approaches to the tasks, along with results from preliminary experiments supporting our ideas. The study is wrapped up and future work is presented in section 6.

2 Autism and simile comprehension

2.1 Autism and figurative language

Highly abstract or figurative metaphors and similes may be problematic for certain groups of language users amongst which are people with different types of acquired language disorders (aphasias) or developmental ones like ASD. As a result of impairment in communication, social interaction and behaviour, ASD are characterised by atypical information processing in diverse areas of cognition (Skoyles, 2011). People with autism, especially if they are children, experience disturbing confusion when confronted with figurative language. Happé (1995) describes:

> A request to "Stick your coat down over there" is met by a serious request for glue. Ask if she will "give you a hand", and she will answer that she needs to keep both hands and cannot cut one off to give to you. Tell him that his sister is "crying her eyes out" and he will look anxiously on the floor for her eye-balls...

The decreased ability of autistic people to understand metaphors and figurative language as a whole (Rundblad and Annaz, 2010; MacKay and Shaw, 2004; Happé, 1995), could be seen as an obstacle in communication, given that we all "think in metaphors" and a language system is "figurative in its nature" (Lakoff and Johson, 1980). The growing demand to overcome this barrier has led to the investigation of possible ways in which NLP can detect and simplify non-literal expressions in a text.

2.2 Comprehending similes

People with ASD[4] show almost no impairment in comprehending those similes which have literal meaning (Happé, 1995). This relative ease in processing is probably due to the fact that similes contain explicit markers (e.g. *like* and *as*), which evoke comparison between two things in a certain aspect.

With regard to understanding figurative similes, Hobson (2012) describes in the case of fifteen-year-old L.: "He could neither grasp nor formulate similarities, differences or absurdities, nor could he understand metaphor".

Theoretically, one of the most obvious markers of similes, the word *like*, could be a source of a lot of misinterpretations. For example, *like* could be a verb, a noun, or a preposition, depending on the context. Given that autistic people have problems understanding context (Skoyles, 2011), how would an autistic reader perceive the role of *like* in a more elaborate and ambiguous comparison? Another possible linguistic reason for the impaired understanding of similes might be that *like* is used

[2] "metaphor, n." OED Online. June 2004. Oxford University Press. 06 February 2013 http://dictionary.oed.com/

[3] A Google Scholar search for papers containing the word *linguistic* have the word *metaphor* in the title approximately 5000 times, but *simile* only around 645 times. In the ACL anthology, *metaphor* occurs around 1070 times while *simile* occurs 52 times.

[4] With level of cognitive ability corresponding to at least first level of Theory of Mind (Baron-Cohen et al., 1985)

ambiguously in many expressions which are neither similes nor comparisons, such as *I feel like an ice cream* or *I feel like something is wrong*.

Even if the expression does not include such an ambiguous use of *like*, there are other cases in which a person with autism might be misled. For example, if the simile is highly figurative or abstract, it may be completely incomprehensible for people with ASD (e.g. the conventional *Love is like a flame*). A step forward towards the simplification of such expressions is their identification and filtering of the ones that are not problematic. Through manipulations, the difficult aspects such as abstractness, figurativeness, and ambiguity can be attenuated.

3 Relevant literature

Comprehensive theoretical investigations into the expressive power of similes can be found in (Bethlehem, 1996) and (Israel et al., 2004). Weiner (1984) applies ontologies to discriminate simple literal and figurative comparisons (loosely using the term *metaphor* to refer to what we call the intersection of similes and metaphors).

Most of the recent computational linguistics research involving similes comes from Veale. In (Veale and Hao, 2008), the pattern *as* __ *as* __ is exploited to mine salient and stereotypical properties of entities using the Google search engine. A similar process has been applied to both English and Chinese by Li et al. (2012). The *Metaphor Magnet* system presented in (Veale and Li, 2012) supports queries against a rich ontology of metaphorical meanings and affects using the same simple simile patterns. The *Jigsaw Bard* (Veale and Hao, 2011) is a thesaurus driven by figurative conventional similes extracted from the Google Ngram corpus.

The role played by figurative language in the field of text simplification has not been extensively studied outside of a few recent publications (Temnikova, 2012; Štajner et al., 2012).

4 Anatomy of a comparison

4.1 Conventionality: norms and exploitations

The theory of norms and exploitations (Hanks, 2013) describes language norms as "a pattern of ordinary usage in everyday language with which a particular meaning or implicature is associated" and argues that norms can be exploited in different ways in order to "say new things or to say old things in new and interesting ways". This distinction can be applied to similes: *as slow as a snail* is a conventional simile that evokes strong association between slowness and snails. On the contrary, in *she looked like a cross between a Christmas tree and an American footballer* (example adapted from the British National Corpus, henceforth BNC) a person (the topic) is not conventionally associated with a Christmas tree (the vehicle), let alone if it is crossed with a football player. In this example the vehicle is not merely unexpected, it also does not exist as a common pattern, and can, by itself, create amazement.

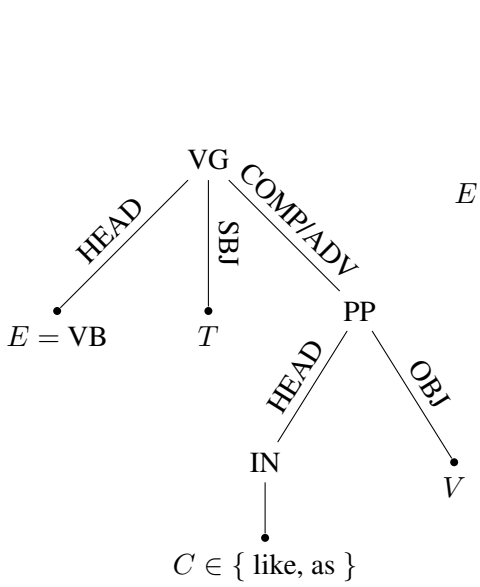
Though figures of speech are good ways to exploit norms, figurative language can become conventional, and an exploitation can be literal (e.g. word creation, ellipsis).

The border between conventionality and creativeness is fuzzy and heuristics such as the ones proposed in (Deignan, 2005) can only approximate it. Possible alternative methods are discussed in section 5.4.2.

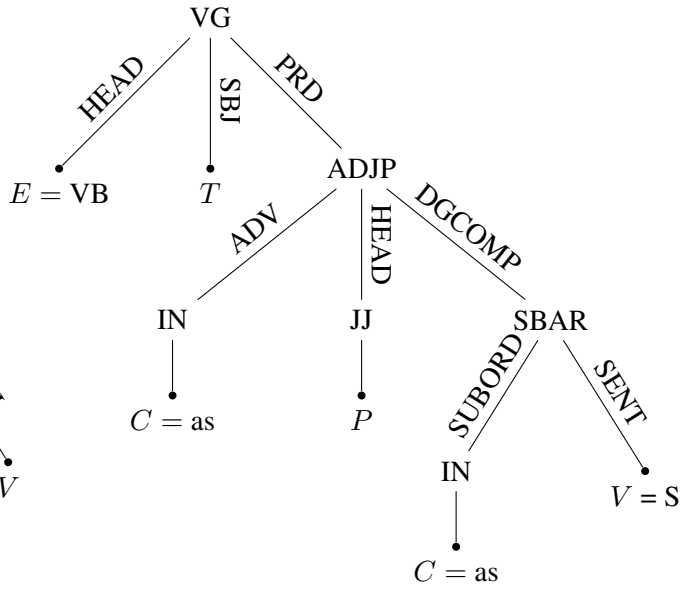
4.2 Syntactic structure

The breadth of comparisons and similes hasn't been extensively studied, so there is no surprise in the small amount of coverage in computational linguistics research on the subject. In order to develop a solid foundation for working with complex comparisons, we will follow and argue for the terminology from (Hanks, 2012), where the structure of a simile is analysed. The same structure applies to comparisons, since as we have said, all similes are comparisons and they are indistinguishable syntactically. The constituents of a comparison are:

- T: the topic, sometimes called tenor: it is usually a noun phrase and acts as logical subject.
- E: the eventuality (event or state): usually a verb, it sets the frame for the observation of the common property.
- P: the shared property or ground: it expresses what the two entities have in common.
- C: the comparator: commonly a preposition (*like* or part of an adjectival phrase (*better than*), it is the trigger word or phrase that marks the presence of a comparison.



(a) Basic comparison pattern. Matches *he eats like a pig* and *it is seen as a release*.



(b) Explicit comparison with double *as*. Matches expressions like *it's as easy as pie*.

Figure 1: GLARF-style representation of two basic comparison patterns.

- V: the vehicle: it is the object of the comparison and is also usually a noun phrase.

An example (adapted from the BNC) of a simile involving all of the above would be:

> [He ^{T}] [looked ^{E}] [like ^{C}] [a broiled frog ^{V}], [hunched ^{P}] over his desk, grinning and satisfied.

The order of the elements is flexible. Fishelov (1993) attributes this reordering to poetic simile, along with other deviations from the norm that he defines as non-poetic simile. We note, in agreement with Bethlehem (1996), that the distinction is rendered less useful when the focus is on the vague notion of poeticality. Fishelov even suggested that poetic similes can be found outside of poetic text, and vice versa. We will therefore focus on exploitations that change the meaning.

More often than not, the property is left for the reader to deduce:

> [His mouth ^{T}] [tasted ^{E}] [like ^{C}] [the bottom of a parrot's cage ^{V}]

But even when all elements appear, the comparison may be ambiguous, as lexical choice in P and in E lead to various degrees of specificity. For example replacing the word *tasted*, which forms the E in the example above, with the more general predicator *is*, results in a simile that might have the same meaning, but is more difficult to decode. On the other hand, the whole V phrase *the bottom of a parrot's cage*, which is an euphemistic metonymy, could be substituted with its concrete, literal meaning thus transforming the creative simile into what might be a conventional pattern. Nested figures of speech can also occur at this level, for example the insertion of a metaphorical and synesthetic P: *it tasted [dirty ^{P}], like a parrot's cage*.

We consider the eventuality E as the syntactic core of the comparison structure. Despite the apparently superior importance of the comparator, which acts as a trigger word, the event acts as a predicator, attracting to it the entire structure in the form of a set of arguments. This observation is missing from the work of Fishelov (1993) and Bethlehem (1996), who lump the event together with either P or T. In terms of meaning, the two constituents are of course tightly connected, but to computationally identify the components, their separation is important.

Roncero (2006) pointed out that for certain common similes (e.g. *love is like a rose*) found on the Internet, it is likely that an explanation of the shared property follows, whereas for all topic-vehicle pairs studied, the corresponding metaphor is less often explained. However, these simpler similes form a special case, as most similes cannot be made into metaphors (Hanks, 2012).

4.3 Comparisons without *like*

Hanks (2012) observes that there are plenty of other ways to make a simile in addition to using *like* or *as*. Most definitions of similes indeed claim that there are more possible comparators, but examples are elusive.

Israel et al. (2004) point out that any construction that can make a comparison can be used to make a simile. This is a crucial point given the amount of flexibility available for such constructions. An example they give is:

> [The retirement of Yves Saint Laurent ^{T}] [is ^{E}] [the fashion equivalent ^{C}] of [the breakup of the Beatles ^{V}]. (heard on the National Public Radio)

We can see that it is possible for the comparator to be informative and not just an empty marker, in this case marking the domain (fashion) to which the topic refers to.

5 Approaches proposed

5.1 Overview

Simplifying creative language involves understanding. The task of understanding similes may be hard to achieve. We will not just write about the components we have already developed (the pattern matching), but also present a broader plan. At a coarse scale, the process breaks down into a syntactic *recognition* step and a semantic step that could be called *entailment*. The goal is to find out what is being said about the topic. Often similes claim that a property is present or absent, but this is not always the case.

5.2 Dataset

At the moment there is no available dataset for comparison and simile recognition and classification. We have begun our investigation and developed the patterns on a toy dataset consisting of the examples from (Hanks, 2005), which are comparisons, similes and other ambiguous uses of the preposition *like* extracted from the BNC. We also evaluated the system on around 500 sentences containing *like* and *as* from the BNC and the VUAMC[5]. The latter features some marking of trigger words, but we chose to score manually in order to assess the relevance of the annotation.

[5]VU Amsterdam Metaphor Corpus (Steen et al., 2010), available at `www.metaphorlab.vu.nl`

5.3 Recognizing comparisons and similes

5.3.1 Comparison pattern matching

We have seen that similes are a subset of comparisons and follow comparison structures. A good consequence is that they follow syntactic patterns that can be recognised. We have used GLARF (Meyers et al., 2001), an argument representation framework built on the output of the BLLIP parser. It enhances the constituency-based parse tree with additional roles and arguments by applying rules and resources like Propbank. The *like* and *as* comparators form the GLARF-style patterns shown in figure 1. The matching process iterates over all nodes with arguments, principally verbs and nominalisations. If the subtree rooted under it matches certain filters, then we assign to the root the role of E and the arguments can fill the other slots.

We evaluated the process on the small development set as well as on the larger set of lexical matches described above. The results are presented in table 1. The mistakes on the development set, as well as many on the other corpus, are caused by slightly different patterns (e.g. *he didn't look much like a doctor*). This can be addressed by adjustment or through automatic discovery of patterns. Expressions like in *hold your hands like this* are mistaken as comparisons. Ad hoc set constructions are mostly correctly unmatched (e.g. *big earners like doctors and airline pilots* but incorrectly matches semantically ambiguous uses of *feel like*.

On the lexical matches of *as*, the behaviour is different as the word seems much less likely to be a trigger. Most errors are therefore returning spurious matches, as opposed to *like*, where most errors are omissions This suggests that each trigger word behaves differently, and therefore robustness across patterns is important.

Overall, our method handles typical comparisons in short sentences rather well. Complex or long sentences sometimes cause T and V to be incompletely identified, or sometimes the parse to fail. This suggests that deep syntactic parsing is a limitation of the approach.

5.3.2 Discovering new patterns

Using a seed-based semi-supervised iterative process, we plan to identify most of the frequent structures used to build conventional comparisons. We expect that, in addition to idiomatic expressions, some T-V pairs often compared to each

	full	part	none
comparison	24	5	4
not comparison	1	1	5

(a) Counts of 40 examples with *like* from the development set in (Hanks, 2005). Partial match $P = 94\%$, $R = 88\%$.

	full	part	none
comparison	0.17	0.07	0.33
not comparison	0.05	0.05	0.33

(b) Proportions of 410 examples with *like* from BNC and VUAMC. Partial match $P = 70.5\%$, $R = 41.7\%$

	full	part	none
comparison	0.11	0.05	0.09
not comparison	0.26	0.11	0.39

(c) Proportions of 376 examples with *as* from BNC and VUAMC. Partial match $P = 29.6\%$, $R = 64.8\%$

Table 1: Confusion matrices and precision/recall scores for comparison identification. Full matching is when the heads of T, E, V and C are correctly identified, while partial is if only some of them are.

other with the *like* pattern will occur in other syntactical patterns or lexical collocations.

5.4 Semantic aspects

5.4.1 Classifying comparisons

The phrases that match patterns like the ones described are not necessarily comparisons. Due to ambiguities, sentences such as *I feel like an ice cream* are indistinguishable from comparisons in our model.

Another aspect we would like to distinguish is whether an instance of a pattern is a simile or not. We plan to tackle this using machine learning. Semantic features from an ontology like the one used in PDEV[6], or a more comprehensive work such as WordNet[7], can carry the information whether T and V belong to similar semantic categories. We expect other information, such as distributional and distributed word vector representations, to be of use.

5.4.2 Conventional similes

It may also be of interest to decide whether an instance is conventional or creative. This can be implemented by measuring corpus frequencies. Instead of looking for perfect matches, patterns can be applied to simply count how many times something is compared to a V, regardless of the specific syntax used[8].

5.4.3 Simplification

The goal of text simplification is to generate syntactically well-formed language[9] that is easier to understand than the original phrase.

A comparison can be formalized as predicate $E(T; P)$. We can think of *his mouth tasted like the bottom of a parrot's cage* as a way to express **taste(his mouth; very bad)**. There is more than one way to build such an encoding.

The task reduces to the generation a simple phrase of the form $T'E'P'$, by simplifying the elements of the representation above. Useful resources are corpus occurrence counts of related phrases, word similarity and relatedness, and conventional associations.

6 Conclusions and future work

The problem of automatic identification of similes has its place in the paradigm of text simplification for people with language impairments. In particular, people with ASD have difficulties understanding figurative language.

We applied the idea of comparison patterns to match subtrees of an enhanced parse tree to easily match comparison structures and their constituents. This lead us to investigate corpus-driven mining of new comparison patterns, to go beyond *like* and *as*.

We are working on semi-automatically developing a dataset of comparisons and ambiguous non-comparisons, labelled with the interesting properties and with a focus on pattern variety and ambiguous cases. This will be useful for evaluating our system at a proper scale. We plan to perform extrinsic evaluation with respect to tasks like text simplification, textual entailment and machine translation.

Acknowledgements

The research described in this paper was partially funded by the European Commission through the FIRST project (FP7-287607) and partially by the BCROCE project.

[6] http://deb.fi.muni.cz/pdev/

[7] http://wordnet.princeton.edu/

[8] Care must be taken to avoid contradictions from exploitations: *The aircraft is like a rock* or *is built like a rock* seems like a conventional simile, but *The aircraft would gently skip like a rock and then settle down on the surface of the ocean* (Example from the BNC) is unconventional.

[9] Especially for ASD readers, who are very sensitive to language mistakes to the point that it completely distracts them from the meaning.

References

Aristoteles and Lane Cooper. 1932. *The rhetoric of Aristotle*. Appleton.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46.

Louise Shabat Bethlehem. 1996. Simile and figurative language. *Poetics Today*, v17(n2):p203(38). table.

Alice Deignan. 2005. *Metaphor and Corpus Linguistics*. Converging Evidence in Language and Communication Research Series. John Benjamins.

David Fishelov. 1993. Poetic and non-poetic simile: Structure, semantics, rhetoric. *Poetics Today*, pages 1–23.

Patrick Hanks. 2005. Similes and Sets: The English Preposition 'like'. In R. Blatná and V. Petkevic, editors, *Languages and Linguistics: Festschrift for Fr. Cermak*.

Patrick Hanks. 2012. The Roles and Structure of Comparisons, Similes, and Metaphors in Natural Language (An Analogical System). In *Presented at the Stockholm Metaphor Festival*, September 6-8.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. Mit Press.

Francesca G. E. Happé. 1995. Understanding minds and metaphors: Insights from the study of figurative language in autism. *Metaphor and Symbolic Activity*, 10(4):275–295.

R. Peter Hobson. 2012. Autism, literal language and concrete thinking: Some developmental considerations. *Metaphor and Symbol*, 27(1):4–21.

Michael Israel, Jennifer Riddle Harding, and Vera Tobin. 2004. On simile. *Language, Culture, and Mind. CSLI Publications*.

George Lakoff and Mark Johson. 1980. *Metaphors We Live By*. Cátedra.

Bin Li, Jiajun Chen, and Yingjie Zhang. 2012. Web based collection and comparison of cognitive properties in english and chinese. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 31–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gilbert MacKay and Adrienne Shaw. 2004. A comparative study of figurative language in children with autistic spectrum disorders. *Child Language Teaching and Therapy*, 20(1):13–32.

Adam Meyers, Ralph Grishman, Michiko Kosaka, and Shubin Zhao. 2001. Covering treebanks with glarf. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources - Volume 15*, STAR '01, pages 51–58, Stroudsburg, PA, USA. Association for Computational Linguistics.

Carlos Roncero, John M. Kennedy, and Ron Smyth. 2006. Similes on the internet have explanations. *Psychonomic Bulletin & Review*, 13:74–77.

Gabriella Rundblad and Dagmara Annaz. 2010. The atypical development of metaphor and metonymy comprehension in children with autism. *Autism*, 14(1):29–46.

John R Skoyles. 2011. Autism, context/noncontext information processing, and atypical development. *Autism research and treatment*, 2011.

G.J. Steen, A.G. Dorst, and J.B. Herrmann. 2010. *A Method for Linguistic Metaphor Identification: From Mip to Mipvu*. Converging evidence in language and communication research. Benjamins.

Irina Temnikova. 2012. *Text Complexity and Text Simplification in the Crisis Management domain*. Ph.D. thesis, University of Wolverhampton, Wolverhampton, UK, May.

Tony Veale and Yanfen Hao. 2008. A context-sensitive framework for lexical ontologies. *Knowledge Eng. Review*, 23(1):101–115.

Tony Veale and Yanfen Hao. 2011. Exploiting readymades in linguistic creativity: A system demonstration of the jigsaw bard. In *ACL (System Demonstrations)*, pages 14–19. The Association for Computer Linguistics.

Tony Veale and Guofu Li. 2012. Specifying viewpoint and information need with affective metaphors: A system demonstration of the metaphor-magnet web app/service. In *ACL (System Demonstrations)*, pages 7–12. The Association for Computer Linguistics.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In Luz Rello and Horacio Saggion, editors, *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, page 14, Istanbul, Turkey, may. European Language Resources Association (ELRA).

E. Judith Weiner. 1984. A knowledge representation approach to understanding metaphors. *Comput. Linguist.*, 10(1):1–14, January.