Semantic Word Sketches

Diana McCarthy
Theoretical and Applied Linguistics
Univ. Cambridge, UK
diana@dianamccarthy.co.uk

Adam Kilgarriff
Lexical Computing Ltd.
Brighton, UK
adam.kilgarriff@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing Ltd.
Masaryk University, Czech Republic
milos.jakubicek@sketchengine.co.uk

Siva Reddy
Institute for Language, Cognition and Computation,
Univ. Edinburgh, UK
siva.reddy@ed.ac.uk

A central task of linguistic description is to identify the semantic and syntactic profiles of the words of a language: what arguments (if any) does a word (most often, a verb) take, what syntactic roles do they fill, and what kinds of arguments are they from a semantic point of view: what, in other terminologies, are their selectional restrictions or semantic preferences. Lexicographers have long done this 'by hand'; since the advent of corpus methods in computational linguistics it has been an ambition of computational linguists to do it automatically, in a corpus-driven way, see for example (Briscoe et al 1991; Resnik 1993; McCarthy and Carroll 2003; Erk 2007).

In this work we start from word sketches (Kilgarriff et al 2004), which are corpus-based accounts of a word's grammatical and collocational behaviour. We combine the techniques we use to create these word sketches with a 315-million-word subset of the UKWaC corpus which has been automatically processed by SuperSense Tagger (SST)[1] (Ciaramita and Altun 2006) to annotate all content words with not only their part-of-speech and lemma, but also their WordNet (Fellbaum 1998) lexicographer class. WordNet lexicographer classes are a set of 41 broad semantic classes that are used for organizing the lexicographers work. These semantic categories group together the WordNet senses (synsets) and have therefore been dubbed `supersenses' (Ciaramita and Johnson, 2003). There are 26 such supersenses for nouns and 15 for verbs. Table 1 provides a few examples and the full set can be seen in Ciaramita and Altun, (2006). SST performs coarse word sense disambiguation, to identify which WordNet supersense a word belongs to. We note that this is not a case where WSD accuracy is critical. In the spirit of Kilgarriff (1997), it is 'background' WSD used for developing a lexical resource. It is hoped that, to a large extent, individual errors in disambiguation are filtered out as noise by the signal from the correct cases.

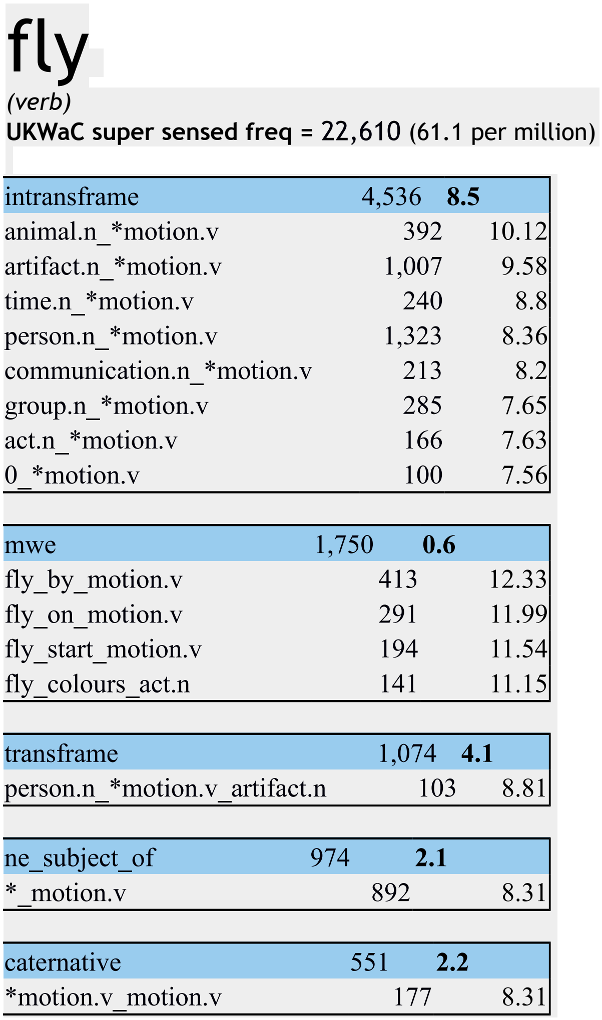
Noun Supersense	**Nouns denoting**
act	acts or actions
animal	animals
artifact	man-made objects
...	
Verb Supersense	**Verbs of**
body	grooming, dressing and bodily care
consumption	eating and drinking
communication	telling, asking, ordering, singing
...	….

Table 1: Some example supersense labels and short descriptions from the WordNet documentation

Each entry in a word sketch shows a different combination of supersenses within a syntactic relationship. As an example, Figure 1 shows a semantic word sketch for the English verb *fly.* The semantic word sketches are produced by a bespoke 'sketch grammar' which identifies the grammatical and collocational behaviour using the part-of-speech tags and finds the predominant semantic classes of the arguments in the syntactic relationships using the supersenses associated with the words in the corpus.

The sketch is presented in tables where the column header states the syntactic pattern identified (e.g. *intransframe*). Then, within each pattern, the head arguments are indicated by the supersense labels (with a part-of-speech suffix) with an asterisk indicating the supersense of the target word in each case (***motion** in these examples for *fly*) .

[1] SST is available at http://sourceforge.net/projects/supersensetag/



fly
(verb)
UKWaC super sensed freq = 22,610 (61.1 per million)

intransframe	4,536	**8.5**
animal.n_*motion.v	392	10.12
artifact.n_*motion.v	1,007	9.58
time.n_*motion.v	240	8.8
person.n_*motion.v	1,323	8.36
communication.n_*motion.v	213	8.2
group.n_*motion.v	285	7.65
act.n_*motion.v	166	7.63
0_*motion.v	100	7.56

mwe	1,750	**0.6**
fly_by_motion.v	413	12.33
fly_on_motion.v	291	11.99
fly_start_motion.v	194	11.54
fly_colours_act.n	141	11.15

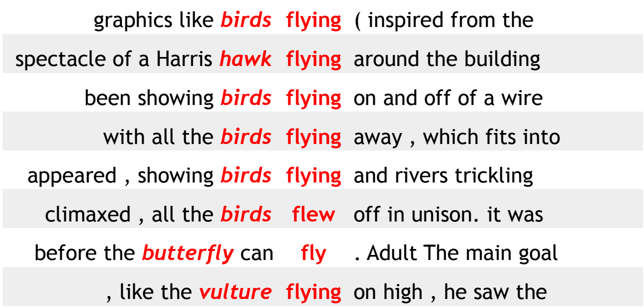
transframe	1,074	**4.1**
person.n_*motion.v_artifact.n	103	8.81

ne_subject_of	974	**2.1**
*_motion.v	892	8.31

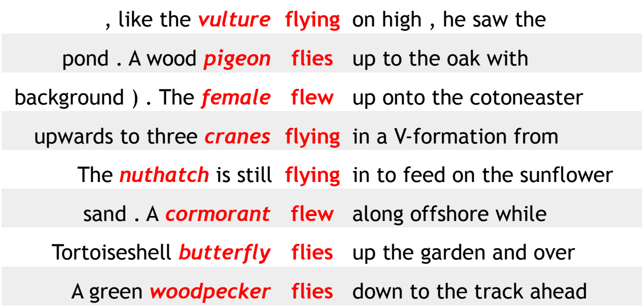
caternative	551	**2.2**
*motion.v_motion.v	177	8.31

Figure 1: Semantic word sketch for English *fly.*

The first and most salient table, *intransframe* lists intransitive frames with the first being **animal.n_*motion.v** – where the verb has an animal subject. There were 392 hits, with a logdice[2] salience score of 10.12. Clicking on the number, we can explore the concordances, as shown in Figure 2. As can be seen, these are valid instances of this frame.



graphics like ***birds***	**flying**	(inspired from the
spectacle of a Harris ***hawk***	**flying**	around the building
been showing ***birds***	**flying**	on and off of a wire
with all the ***birds***	**flying**	away , which fits into
appeared , showing ***birds***	**flying**	and rivers trickling
climaxed , all the ***birds***	**flew**	off in unison. it was
before the ***butterfly*** can	**fly**	. Adult The main goal
, like the ***vulture***	**flying**	on high , he saw the



, like the ***vulture***	**flying**	on high , he saw the
pond . A wood ***pigeon***	**flies**	up to the oak with
background) . The ***female***	**flew**	up onto the cotoneaster
upwards to three ***cranes***	**flying**	in a V-formation from
The ***nuthatch*** is still	**flying**	in to feed on the sunflower
sand . A ***cormorant***	**flew**	along offshore while
Tortoiseshell ***butterfly***	**flies**	up the garden and over
A green ***woodpecker***	**flies**	down to the track ahead

Figure 2: concordance (truncated to fit) for **animal.n** as subject of *fly.v.*

Next comes the **artifact.n_*motion.v** i.e. artifact-as-subject frame, top lemmas in the subject slot being *plane, flag, ball, aircraft, helicopter, shot, airline, bullet.* It is not immediately apparent if this is a separate sense of *fly* to animal-as-subject: it depends on whether the user (such as a lexicographer) making the choice is a 'lumper' or a 'splitter': how fine-grained they like their senses to be.[3] It is also not clear whether *plane* (which is self-moving) fits in the same sense as *ball* (which is not; or *flag,* possibly an intermediate case).

time.n_*motion.v relates to the idiom of time flying past, with lemmas in addition to *time* itself being *day, night, hour, year* and *winter.*

The next table, **mwe** (multiword expresssions), covers two prepositional verbs and two idioms: *a flying start* and *flying colours,* which cannot usefully (from a semantic, or lexicographic, point of view) be treated elsewhere in the entry.

The third table, **person.n_*motion.v_artifact.n**, for transitives, here has just one frame (over the frequency threshold; here set at 100). These are mostly people flying aircraft, with occasional instances, missed by the *mwe* filter, of people flying the flag.

The fourth table and its single frame cover cases where named entities (*ne*'s – named people and organisations) are doing the flying. The fifth and final one covers some problematic parsing cases: *fly tipping* (a British idiom meaning depositing rubbish illegally), *fly casting* and *fly reel* (technical terms from the fishing domain) and *flying trapeze* (from the circus).

Formalism

The formalism in which the sketch grammar (which specifies the word sketch) is written is an extended and augmented version of the one described in Kilgarriff et al. (2004), itself adopted and adapted from Christ and Schulze (1994). The

[2] See Rychlý 2008.

[3] They may also be working to other constraints, like never giving the same sense where a key argument has a different lexicographer class.

full documentation is available at the Sketch Engine website.[4]

Relation to FrameNet and similar projects and to distributional semantics

FrameNet (Baker et al. 1998) has been probably the most ambitious and inspiring project for building a lexical resource in recent years. It aims to establish the set of semantic frames for the predicates of English, complete with a description of the semantic roles of the arguments (the frame elements) in each frame and their syntactic and semantic specifications. It has been a highly influential project, spawning a wide variety of subsequent projects, for example to produce FrameNets for other languages, to automate the process of building FrameNet-like resources, and to disambiguate words according to frames.

FrameNet is a manual project: people decide what the frames are, what instances fit which frames, and so forth, albeit with sophisticated computational support. This assures high accuracy, but also slow progress. The contribution that semantic word sketches might make to FrameNet-style projects (including Corpus Pattern Analysis, (CPA: Hanks 2013) and VerbNet (Kipper et al 2006)) could be

- helping in extending the coverage of the dataset by providing data to explore, edit and include
- providing an additional type of data that is not currently available within FrameNet: the actual argument slot fillers for a given frame with frequency data and corpus examples.
- research into what parts of, and to what extent, the FrameNet-style lexicography process can be automated, where manual entries provide a gold standard which semantic word sketches aspire to. One important aspect not covered by semantic word sketches is the role of the semantic arguments within a frame (for example whether the subject of *fly* is riding a vehicle or a self-mover).

Semantic word sketches offer the benefits that the corpora, WordNets (or their younger relation, Babelnet (Navigli and Ponzetto 2012)) and the computational tools needed to create them are already in place for a number of languages, so the investment needed to create them for a new language is modest.

Semantic word sketches contrast with FrameNet, VerbNet and CPA through being automatic. There is another stream of work with similar goals but that is completely data-driven and, unlike semantic word sketches, does not use a manually created resource (WordNet) for defining semantic classes. This stream is distributional semantics (see Baroni and Lenci (2010) for an overviewand Socher (2014) on the recent and related area of 'deep learning'). We look forward to exploring the contrasts and complementarities between semantic word sketches and distributional semantics.

[4] http://www.sketchengine.co.uk

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The Berkeley FrameNet Project. *Proc. ACL.*

M. Baroni and A. Lenci. 2010. Distributional Memory: A general framework for corpus-based semantics. Computational Linguistics 36(4): 673-721.

Ciaramita, M. and Altun, Y. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. Proc EMNLP, Sydney, Australia: pp 594-602.

Ciaramita M. and Johnson, M 2003. Supersense Tagging of Unknown Nouns in WordNet. In Proceedings of EMNLP 2003.

Erk K. 2007. A simple, similarity-based model for selectional preferences. Proc. ACL 2007. Prague, Czech Republic, 2007.

Fellbaum, C editor. 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge.

Hanks, P. W. 2013. Lexical Analysis: a theory of norms and exploitations. MIT Press.

Kilgarriff, A. 1997. Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction. Proc Workshop on Lexicon-driven Information Extraction, Frascati, Italy.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. Proc. EURALEX. pp. 105–116.

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. 2006. Extending VerbNet with novel verb classes. *Proc. LREC*.

McCarthy, D. and Carroll J. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences, *Computational Linguistics,* 29(4). pp 639-654.

Navigli. R., and S. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217-250.

Resnik. P. 1993. Selection and Information: A Class-Based Approach to Lexical Relationships. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Rychlý, P. 2008. A Lexicographer-Friendly Association

Score. Proc. RASLAN workshop, Brno, Czech Republic.

Schulze, B. M., & Christ, O. 1994. The CQP user's manual. *Universität Stuttgart*, *Stuttgart*.

Socher, R. 2014. Recursive Deep Learning for Natural Language Processing and Computer Vision, PhD Thesis, Computer Science Department, Stanford University