

## Semantic Word Sketches

Diana McCarthy,<sup>†</sup> Adam Kilgarriff,<sup>◇</sup>  
Miloš Jakubíček,<sup>◇‡</sup> Siva Reddy<sup>\*</sup>

DTAL University of Cambridge<sup>†</sup>, Lexical Computing<sup>◇</sup>,  
University of Edinburgh<sup>\*</sup>, Masaryk University<sup>‡</sup>

July 2015

# Outline

- 1 The Sketch Engine
  - Concordances
  - Word Sketches
- 2 Semantic Tagging
  - Super Sense Tagger (SST)
  - SST Supersenses
- 3 Semantic Tags in Sketch Engine
  - In the Concordance
  - Semantic Word Sketches
  - Other Possibilities from SST Output
- 4 Comparison to FrameNet
- 5 Conclusions

# The Sketch Engine

- concordances, word lists, collocations
- word sketches
  - create and examine syntactic profiles and collocations of words
  - input automatic part-of-speech tags and a bespoke 'sketch grammar'
- automatic thesauruses: which other words have similar profiles?
- sketch differences between words

# The Sketch Engine

for viewing corpora

Concordance  
Word List  
Word Sketch  
Thesaurus  
Sketch-Diff  
Sketch-Eval  
Corpus Info  
Manage corpus  
My jobs

Save  
as subcorpus  
View options  
KWIC  
Sentence  
Sort  
Left  
Right

Query **mouse** 49,985 > Random sample 250 (0.16 per million)

Page  of 13  [Next](#) | [Last](#)

<a href="#">site=sange...</a>	completed the first phase of reading the <b>mouse</b> 'book of life', reaching its goal on time
<a href="#">site=i-sis...</a>	supplements may reverse the damage. Obese <b>mice</b> given dietary supplements were found to
<a href="#">site=sange...</a>	identity). Thus, by comparing human and <b>mouse</b> genome sequences, the regions of high similarity
<a href="#">site=sange...</a>	set of BAC clones that covers the entire <b>mouse</b> genome is being sequenced. The BAC data
<a href="#">site=phgu...</a>	extraembryonic stem cell lines derived from single <b>mouse</b> blastomeres. Nature advance online publication
<a href="#">site=sange...</a>	investigations of the function of genes using the <b>mouse</b> as model genetic system. The principal
<a href="#">site=birdt...</a>	all the group members who were as quiet as <b>mice</b> . We eventually heard Grey-headed Quail-dove
<a href="#">site=manch...</a>	their experiments on monkey kidney cells and <b>mouse</b> skin cells with similar results. Email
<a href="#">site=littl...</a>	very much. To be honest, I'd be fine with <b>mice</b> if they were in a cage. Mice are pretty
<a href="#">site=blog...</a>	said that it wasnt for me it was for my <b>mouse</b> as way of an explanation - she just raised
<a href="#">site=i-sis...</a>	whereas those from Sertoli cells of immature <b>mice</b> died at an unusually early age. By contrast
<a href="#">site=maryw...</a>	the same graphical interaction without a <b>mouse</b> or other pointing device. Alternative text
<a href="#">site=www-e...</a>	tissue reactions in normal and immunized <b>mice</b> to a reticulotropic strain of Trypanosoma
<a href="#">site=theno...</a>	open in a new window Open Map ROLL your <b>mouse</b> on and off the EDGE of the maps to reveal
<a href="#">site=readg...</a>	work with Plasmodium chabaudi in laboratory <b>mice</b> (Buckling, Taylor et al. 1997; Taylor, Walliker
<a href="#">site=pract...</a>	particularly quirky, and involved using the word " <b>mouse</b> " wherever possible inside words, as well

# The Sketch Engine

## Word Sketches: syntactic profiles

Sketch Engine Send feedback corpus: ukWaC

Concordance  
 Word List  
 Word Sketch  
 Thesaurus  
 Sketch-Diff  
 Sketch-Eval  
 Corpus Info  
 Manage corpus  
 My jobs  
 ?

---

Save  
 Change options  
 Clustering  
 Sorting

**eat** *(verb)*  
 ukWaC freq = [137,789](#) (88.34 per million)

<a href="#">object</a>	<a href="#">60,826</a>	<a href="#">5.50</a>	<a href="#">subject</a>	<a href="#">16,966</a>	<a href="#">2.20</a>	<a href="#">modifier</a>	<a href="#">22,796</a>	<a href="#">0.70</a>	<a href="#">and/or</a>	<a href="#">9,742</a>	<a href="#">0.70</a>
meat	<a href="#">1,802</a>	8.71	binge	<a href="#">92</a>	7.21	healthily	<a href="#">662</a>	9.82	drink	<a href="#">3,168</a>	10.59
disorder	<a href="#">1,915</a>	8.57	LTY	<a href="#">63</a>	6.81	properly	<a href="#">327</a>	6.84	sleep	<a href="#">681</a>	8.31
meal	<a href="#">1,812</a>	8.17	RECOMMEND	<a href="#">47</a>	6.47	sensibly	<a href="#">90</a>	6.82	cook	<a href="#">214</a>	7.04
food	<a href="#">5,233</a>	8.16	ye	<a href="#">89</a>	5.84	enough	<a href="#">212</a>	6.35	shop	<a href="#">73</a>	6.82
diet	<a href="#">1,376</a>	8.05	lion	<a href="#">68</a>	5.83	away	<a href="#">437</a>	6.35	socialise	<a href="#">45</a>	6.58
fruit	<a href="#">1,298</a>	7.84	vegetarian	<a href="#">40</a>	5.78	regularly	<a href="#">225</a>	6.23	breathe	<a href="#">87</a>	6.41
lunch	<a href="#">1,070</a>	7.74	flesh	<a href="#">77</a>	5.73	alfresco	<a href="#">47</a>	6.05	swallow	<a href="#">57</a>	6.36
breakfast	<a href="#">910</a>	7.72	slug	<a href="#">37</a>	5.69	outdoors	<a href="#">50</a>	6.01	relax	<a href="#">133</a>	5.97
habit	<a href="#">751</a>	7.60	caterpillar	<a href="#">34</a>	5.69	happily	<a href="#">66</a>	5.90	smoke	<a href="#">56</a>	5.81
bread	<a href="#">694</a>	7.53	cheap	<a href="#">30</a>	5.67	together	<a href="#">455</a>	5.88	chew	<a href="#">31</a>	5.74

# Sketch Grammars

## Under the hood

- Definitions: `define('any_noun', "N..")`

...

- Relations

`=subject/subject_of`

`2:any_noun rel_start? adv_aux_string_incl_be 1:verb_not_pp`

`2:any_noun rel_start? adv_aux_string_incl_be aux_have adv_string 1:past_part`

`1:past_part adv_string [word="by"] long_np`

## Semantic Class Tagging

- aim to build word sketches on syntactic and semantic information
- automatic 'superclass' tagging technology
- superclass: a coarse grained semantic class that is applicable to multiple words (e.g. **animal** for *cat*, *fly*, *hare*, *pig* etc. . .
- allow search and analysis with these classes and
- semantic word sketches: basic semantic frame with semantic preferences for arguments

## Semantic Class Tagging

Super Sense Tagger (SST) Ciaramita and Altun (2006)  
(<http://sourceforge.net/projects/supersensetag/>)

- semantic tags are WordNet Fellbaum (1998) lexicographer classes
- supervised word sense disambiguation (i.e. it requires hand labelled data for training) using a Hidden Markov Model  
e.g. labels *mouse* as **animal**, **artifact**)
- SemCor (Landes et al., 1998) used as training data
- Named Entity Recognition  
e.g. < *RHM Technology Ltd.* > **organization**
- Multiword tagging using multiwords from WordNet  
e.g. *couch potato*



## SST WordNet Noun Classes (25)

act	acts or actions
object	natural objects (not man-made)
animal	animals
quantity	quantities and units of measure
artifact	man-made objects
phenomenon	natural phenomena
attribute	attributes of people and objects plant plants
food	food and drinks
...	...

## SST WordNet Verb Classes (15)

body	grooming, dressing and bodily care
emotion	feeling
change	size, temperature change, intensifying
motion	walking, flying, swimming
cognition	thinking, judging, analyzing, doubting
perception	seeing, hearing, feeling
communication	telling, asking, ordering, singing
possession	buying, selling, owning
creation	sewing, baking, painting, performing
...	...

## Experiments

- just over 25% of the UKWaC Ferraresi et al. (2008)
- SST tagged with
  - part-of-speech tags (Penn TreeBank)
  - supersenses (WordNet labels)
  - Named Entity Labels
  - WordNet multiwords

# Semantic Tags in the Concordance

Corpus: UKWaC super sensed

Hits: 15772 (42.6 per million)

[First](#) | [Previous](#) Page  of 789  [Next](#) | [Last](#)

<a href="#">#1493049</a>	credibility that surrounds the <mwe><ne> Mickey	Mouse	/NNP/other.n	</ne></mwe><m
<a href="#">#1496976</a>	hovering , <mwe> seeking out </mwe> a vole or	mouse	/NN/animal.n	. INFORMATION F
<a href="#">#1545720</a>	audio visuals and the <mwe><ne> Soviet Spy	Mouse	/NNP/other.n	Trail </ne></mw
<a href="#">#1561637</a>	even after disruption . Control through the	mouse	/NN/animal.n	Interactive stori
<a href="#">#1561653</a>	keyboard . They are controlled entirely by the	mouse	/NN/animal.n	which moves the
<a href="#">#1561673</a>	of effects . In a very real sense , the	mouse	/NN/animal.n	represents contr
<a href="#">#1561693</a>	importance of access to the controlling device (	mouse	/NN/animal.n	or keyboard , <m
<a href="#">#1561728</a>	</mwe> , 1993 ) . Seen in this light , the	mouse	/NN/animal.n	might be consid
<a href="#">#1561813</a>	different methods were used when passing the	mouse	/NN/animal.n	to the <ne> next
<a href="#">#1561826</a>	Sometimes , <mwe> for example </mwe> , when the	mouse	/NN/animal.n	had been left in
<a href="#">#1561862</a>	remaining member of the group moved the	mouse	/NN/animal.n	towards the nex
<a href="#">#1561928</a>	within the group . If simply leaving the	mouse	/NN/animal.n	is seen as unhel
<a href="#">#1562003</a>	members who were not in possession of the	mouse	/NN/animal.n	issued a significa
<a href="#">#1562048</a>	in <mwe> actual possession </mwe> . While the	mouse	/NN/animal.n	gave the undispr
<a href="#">#1562081</a>	directed the same commands to the holder of the	mouse	/NN/animal.n	, adding extra ps
<a href="#">#1608486</a>	and dragging the control points using the	mouse	/NN/animal.n	( left button ) or
<a href="#">#1608563</a>	dialog or dragging the points using left <mwe>	mouse	/NN/artifact.n	button </mwe> .
<a href="#">#1664217</a>	with a central finger-ball ( e.g. marble	mouse	/NN/animal.n	) . Just more phy
<a href="#">#1768071</a>	students attending workshops ( eg. <mwe>	mouse	/NN/artifact.n	mats </mwe> , p
<a href="#">#1831217</a>	</ne></mwe> include the <mwe><ne> Yellow Necked	Mouse	/NNP/other.n	</ne></mwe> , V

[First](#) | [Previous](#) Page  of 789  [Next](#) | [Last](#)

# Semantic Tags in the Word Sketch (selected)

**eat** (*verb*) UKWaC super sensed freq = 26329 (71.2 per million)

<u>transframe</u>	<u>1241</u>	<u>6.6</u>	<u>Intransframe</u>	<u>1021</u>	<u>2.7</u>
person.n_*consumption.v_food.n	<u>178</u>	11.42	animal.n_*consumption.v	<u>85</u>	10.84
group.n_*consumption.v_food.n	<u>57</u>	10.25	person.n_*consumption.v	<u>382</u>	10.25
person.n_*consumption.v_plant.n	<u>37</u>	9.77	group.n_*consumption.v	<u>145</u>	9.82
person.n_*consumption.v_animal.n	<u>35</u>	9.62	0_*consumption.v	<u>61</u>	9.73
animal.n_*consumption.v_animal.n	<u>30</u>	9.56	state.n_*consumption.v	<u>20</u>	8.89
animal.n_*consumption.v_plant.n	<u>25</u>	9.32	time.n_*consumption.v	<u>19</u>	8.74
animal.n_*consumption.v_food.n	<u>24</u>	9.21	communication.n_*consumption.v	<u>30</u>	8.57
person.n_*consumption.v_person.n	<u>52</u>	8.97	artifact.n_*consumption.v	<u>30</u>	8.51
0_*consumption.v_food.n	<u>20</u>	8.92	food.n_*consumption.v	<u>19</u>	8.4
animal.n_*consumption.v_artifact.n	<u>19</u>	8.79	other.n_*consumption.v	<u>15</u>	8.1

# Semantic Tags in the Word Sketch (selected)

**laugh** (*verb*) UKWaC super sensed freq = 6489 (17.5 per million)

<u>V_PP</u>	<u>148</u>	<u>9.1</u>	<u>Intransframe</u>	<u>1101</u>	<u>8.9</u>
*body.v_at_cognition.n	<u>14</u>	11.02	person.n_*body.v	<u>556</u>	10.49
*body.v_at_communication.n	<u>12</u>	10.94	group.n_*body.v	<u>143</u>	10.02
*body.v_at_person.n	<u>7</u>	9.77	0_*body.v	<u>102</u>	10.33
*body.v_as_person.n	<u>6</u>	9.78	artifact.n_*body.v	<u>49</u>	9.19
*communication.v_at_location.n	<u>6</u>	7.8	time.n_*body.v	<u>23</u>	8.49
*body.v_in_cognition.n	<u>5</u>	9.58	location.n_*body.v	<u>16</u>	7.88
*communication.v_at_communication.n	<u>5</u>	8.39	event.n_*body.v	<u>9</u>	7.56
*communication.v_at_person.n	<u>5</u>	8.14	other.n_*body.v	<u>8</u>	7.38
			cognition.n_*body.v	<u>7</u>	7.2
			communication.n_*body.v	<u>7</u>	6.89

# Semantic Word Sketch Grammar

An example for the intransitive frame

=intransframe

\*COLLOC “%(2.sense)\_\*%(1.sense)-x”

2:any\_noun rel\_start? adv\_aux\_string\_incl\_be 1:verb\_not\_pp  
not\_np\_start

2:any\_noun rel\_start? adv\_aux\_string\_incl\_be aux\_have  
adv\_string 1:past\_part not\_np\_start

# MWEs: detected by SST

**bread** (*noun*) UKWaC super sensed freq = 8355 (22.6 per million)

<u>mwe</u>	<u>1887</u>	<u>0.8</u>
bread_and_butter_possession.n	<u>203</u>	11.37
loaf_of_bread_food.n	<u>180</u>	11.25
white_bread_food.n	<u>171</u>	11.3
garlic_bread_food.n	<u>74</u>	10.22
brown_bread_food.n	<u>71</u>	10.16
french_bread_food.n	<u>53</u>	9.77
unleavened_bread_food.n	<u>42</u>	9.45
bread_street_location.n	<u>34</u>	9.15
bread_maker_person.n	<u>31</u>	9.03
rye_bread_food.n	<u>30</u>	8.98
bread_knife_artifact.n	<u>21</u>	8.48



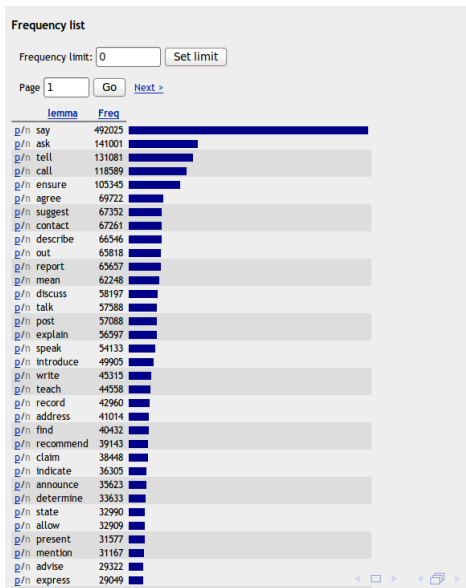
# MWEs: Sketch Diff chip (green) vs chips (red)

mwe	325	314	-2.0	-2.0
fish_and_chip_food.n	0	<u>70</u>	0.0	10.4
tortilla_chip_food.n	0	<u>14</u>	0.0	8.3
potato_chip_food.n	<u>10</u>	<u>40</u>	7.7	9.7
memory_chip_artifact.n	<u>25</u>	<u>32</u>	9.0	9.4
gene_chip_artifact.n	<u>11</u>	<u>10</u>	7.9	7.8
poker_chip_artifact.n	<u>14</u>	<u>12</u>	8.3	8.0
silicon_chip_artifact.n	<u>40</u>	<u>27</u>	9.7	9.1
bargaining_chip_attribute.n	<u>29</u>	<u>7</u>	9.3	7.2
chocolate_chip_food.n	<u>14</u>	0	8.3	0.0
chip_shot_act.n	<u>15</u>	0	8.4	0.0

## Portion of Sketch Diff laugh (green) vs cry (red)

V_PP	148	117	9.1	7.0
*communication.v_in_location.n	0	<u>6</u>	0.0	7.3
*stative.v_for_act.n	0	<u>11</u>	0.0	6.6
*communication.v_at_location.n	<u>6</u>	0	7.8	0.0
*communication.v_at_person.n	<u>5</u>	0	8.1	0.0
*communication.v_at_communication.n	<u>5</u>	0	8.4	0.0
*body.v_in_cognition.n	<u>5</u>	0	9.6	0.0
*body.v_at_person.n	<u>7</u>	0	9.8	0.0
*body.v_as_person.n	<u>6</u>	0	9.8	0.0
*body.v_at_communication.n	<u>12</u>	0	10.9	0.0
*body.v_at_cognition.n	<u>14</u>	0	11.0	0.0

# Semantic Word Lists: CQL + Word Frequency (Communication Verbs)



# Semantic Word Lists: FindX (communication verbs)

0.1	122.9	<a href="#">say-v</a>	<a href="#">271</a>
0.1	119.4	<a href="#">tell-v</a>	<a href="#">62</a>
0.1	112.0	<a href="#">ask-v</a>	<a href="#">75</a>
0.2	101.9	<a href="#">out-v</a>	<a href="#">53</a>
0.2	100.0	<a href="#">humour-v</a>	<a href="#">53</a>
0.2	100.0	<a href="#">critique-v</a>	<a href="#">142</a>
0.2	100.0	<a href="#">underline-v</a>	<a href="#">2166</a>
0.2	100.0	<a href="#">stammer-v</a>	<a href="#">50</a>
0.2	100.0	<a href="#">reintroduce-v</a>	<a href="#">501</a>
0.2	100.0	<a href="#">re-introduce-v</a>	<a href="#">109</a>
0.2	100.0	<a href="#">shriek-v</a>	<a href="#">116</a>
0.2	100.0	<a href="#">exhort-v</a>	<a href="#">88</a>
0.2	100.0	<a href="#">publicise-v</a>	<a href="#">1244</a>
0.2	100.0	<a href="#">chide-v</a>	<a href="#">116</a>
0.3	100.0	<a href="#">interrogate-v</a>	<a href="#">730</a>
0.3	100.0	<a href="#">fate-v</a>	<a href="#">67</a>
0.3	100.0	<a href="#">bemoan-v</a>	<a href="#">277</a>
0.3	100.0	<a href="#">absolve-v</a>	<a href="#">136</a>
0.3	100.0	<a href="#">signpost-v</a>	<a href="#">160</a>
0.3	100.0	<a href="#">unrated-v</a>	<a href="#">321</a>
0.3	100.0	<a href="#">chronicle-v</a>	<a href="#">487</a>
0.3	100.0	<a href="#">telegraph-v</a>	<a href="#">75</a>
0.3	100.0	<a href="#">spam-v</a>	<a href="#">218</a>
0.3	100.0	<a href="#">misquote-v</a>	<a href="#">119</a>
0.4	100.0	<a href="#">extol-v</a>	<a href="#">87</a>
0.4	100.0	<a href="#">eschew-v</a>	<a href="#">322</a>
0.4	100.0	<a href="#">nominates-v</a>	<a href="#">71</a>
0.4	100.0	<a href="#">evince-v</a>	<a href="#">156</a>
0.4	100.0	<a href="#">spooof-v</a>	<a href="#">66</a>
0.4	100.0	<a href="#">rejuvenate-v</a>	<a href="#">205</a>
0.4	100.0	<a href="#">symbolise-v</a>	<a href="#">292</a>
0.4	100.0	<a href="#">pardon-v</a>	<a href="#">445</a>

## Comparing to FrameNet (Ruppenhofer et al., 2010)

- FrameNet contains lots of useful information e.g. [FRAME **employing**:  
Frame Elements: **Employer** **Employee** **Position** **Tasks**  
**Compensation** ...  
Definition: An **Employer** *employs* an **Employee** whose **Position** entails that the **Employee** perform certain **Tasks** in exchange for **Compensation**
- lots of other information
  - lexical units *employ.v* *commision.v* *staff.n* *employment.n*
  - precedes frame **firing**
  - with corpus examples, *I employed him as Chief Gardener for ten years*
- but manually produced so low coverage
- Semantic word sketches can provide additional information and high coverage

## Summary

- semantic tagging alongside part-of-speech for semantic word sketches
- provide syntactic and semantic profiling for
  - semantic queries and word lists
  - semantic and syntactic profiling in the word sketch
  - comparing words by the profiles

## Future Possibilities

- try other semantic tagsets, taggers and tools
- sketch grammar could be developed further
- no identification of semantic roles as yet in contrast to FrameNet (Ruppenhofer et al., 2010), Propbank (Palmer et al., 2005) and VerbNet (Kipper-Schuler, 2005)
- Semantic word sketches could be used to provide selectional preferences and corpus information to such resources

# Thank You

Query **thank, V.\*** 14,920 > Positive filter (excluding KWIC) **you** 5,551 > Shuffle 5,551

Page  of 278  [Next](#) | [Last](#)

<a href="http://www...">http://www...</a>	see , a little more really means a lot -	<b>Thank</b>	/communication.
<a href="http://www...">http://www...</a>	Jamie and Lynne Reilly 29 June , 2002 ``	<b>Thank</b>	/communication.
<a href="http://www...">http://www...</a>	Team , ( Jack , Billy and young Andy , )	<b>THANK</b>	/communication.
<a href="http://www...">http://www...</a>	important member of the family , and we	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	B. I would have liked the opportunity to	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	obtained some very interesting information . I	<b>thank</b>	/communication.f
<a href="http://bee...">http://bee...</a>	again for eternity in heaven . I want to	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	to cover , but thank you for coming and	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	ZENAB ( ) posted : 08.03.2006 message :	<b>Thank</b>	/communication.
<a href="http://www...">http://www...</a>	Joanne.gowing@pizzaexpress.com Happy donating and	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	Platform : GameCube Sent by : Andrew Bernish (	<b>Thank</b>	/communication.
<a href="http://www...">http://www...</a>	to keep Nicholas in all of your prays and	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	top Thank you June 2006 I would like to	<b>thank</b>	/communication.f
<a href="http://spo...">http://spo...</a>	Peter MacLeod . Hello to you too , sir , and	<b>thank</b>	/communication.f
<a href="http://www...">http://www...</a>	, I do . You seem so damn sure . HOMER :	<b>Thank</b>	/communication.



- Ciaramita, M. and Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia. Association for Computational Linguistics.
- Fellbaum, C., editor (1998). *WordNet, An Electronic Lexical Database*. The MIT Press, Cambridge, MA.
- Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.
- Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, Computer and

Information Science Dept., University of Pennsylvania.  
Philadelphia, PA.

- Landes, S., Leacock, C., and Randee, I. T. (1998). Building semantic concordances. In Fellbaum, C., editor, *WordNet: an Electronic Lexical Database*, pages 199–237. MIT Press.
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2010). FrameNet II: Extended theory and practice. Technical report, International Computer Science Institute, Berkeley. <http://framenet.icsi.berkeley.edu/>.