

Verb Behavior is not Verb Nature: Sense and Genre Biases as Sources of Subcategorization Probabilities

Douglas Roland, Daniel Jurafsky, and Laura Michaelis - Department of Linguistics - University of Colorado, Boulder

Introduction

Problem:

Verb subcategorization probability depends on the method of measurement (Merlo 1994, Gibson et al. 1996, Roland & Jurafsky 1997). Does this mean there is no fixed subcategorization probability for a given verb?

Solution:

- **Lemma Argument Probability Hypothesis:** Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.
- **Probabilistic Combination Hypothesis:** Observed Subcategorization Probability = Lemma Argument Probability + Contextual Influence

Methodology

5 Corpora:

- Comminne et al. (1984) (CFJCF) single sentence production
- Garmsey et al. (1997) (Garmsey) single sentence completion
- Brown corpus (BC) Penn Treebank
- Wall Street Journal corpus (WSJ) Penn Treebank
- Switchboard corpus (SWBD) Penn Treebank

166 verbs coded for subcategorization:

- Complementations: FrameNet (Baker et al. 1998).
- 17 major categories: 0, PP, VPro, Sfin, Sfin, VPing, VPbrst, NP, NP NP, [NP PP], [NP Vpro], [NP Swh], [NP Sfin], Quo, Passives, and Other.
- Only true syntactic arguments, no adjuncts, following the distinction made in Treebank (Marcus et al. 1993).
- 7 verbs also hand-coded for Wordnet sense.

The following table shows the sample size for each corpus:

Corpus	Token/Type	examples per verb
CFJCF	14,000 (127 CFJCF verbs)	$n \approx 29, 39, \text{ or } 68$
Garmsey	5,200 (48 Garmsey verbs)	$n \approx 108$
BC	21,000 (127 CFJCF verbs)	$0 \leq n \leq 2,644$
	6,600 (48 Garmsey verbs)	
WSJ	25,000 (127 CFJCF verbs)	$0 \leq n \leq 11,411$
	5,700 (48 Garmsey verbs)	
SWBD	10,000 (127 CFJCF verbs)	$0 \leq n \leq 3,169$
	4,400 (48 Garmsey verbs)	

How do you know if a verb is used the same way in two different corpora?

$$\text{Cosine} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

The cosine of the subcategorization probability vectors for the verb in each corpus can be used as a measure of the degree of difference (Salton & McGill 1983).

	lear	0	PP	Swh	Sfin	VPrst	NP	NP PP	passive
BC	4	12	3	1	15	47	4	14	
WSJ	0	17	3	5	13	56	10	10	

Contextual Influences

Test tube sentences are different from wild sentences.

Why? Connected discourse requires **discourse cohesion**.

Data Source	% passive sentences
Garmsey	0.6%
CFJCF	2.2%
Switchboard	6.7%
Wall Street Journal	7.8%
Brown corpus	7.8%

• **Zero anaphora** used in discourse cohesion - e.g. "The shot reverberated in diminishing whiplashes of sound. *It* hit followed". (BC) (data for verb follow)

Data Source	% [0] subcat frame
Garmsey	0%
CFJCF	5%
Wall Street Journal	11%
Switchboard	2.2%
Brown	

Why? Experimental subjects tend to rely on **default referents**.

• Experimental subjects use more **first person subjects**

Data Source	% first person subject
Garmsey	40%
CFJCF	39%
Switchboard	18%
Brown corpus	7%
Wall Street Journal	

• Experimental subjects use more VP internal NPs that are **anaphorically related to the subject** - e.g. "Alice *prayed* that *her* daughter wouldn't die". (Garmsey)

Data Source	% related subject/NP
Garmsey	41%
CFJCF	26%
Wall Street Journal	15%
Brown corpus	12%
Switchboard	8%

• Experimental subjects use more **prototypical objects** - i.e. 33% of the uses of **accept** had either *award*, *fact*, *job*, or *invitation* as an object in Garmsey. In BC, no common noun occurred in more than 3% of the uses. (data for verb *accept*)

Data Source	token count	type count	Argument token/type ratio
Garmsey	107	54	2.0
CFJCF			
WSJ	138	105	1.3
Brown corpus	112	86	1.3
Switchboard	15	14	1.1

Word-sense Influences

Each verb sense has its own subcategorization probability

• **Corpora have different distributions of verb sense.**

Sense of BC% WSJ% Example of the senses of charge.

attack	23%	0%	<i>His followers ... charged the hill, firing as they ran.</i> (BC)
run	8%	0%	<i>She charged off to the bedrooms.</i> (BC)
appoint	6%	4%	<i>The commission is charged with designing a ... program.</i> (WSJ)
accuse	39%	58%	<i>Separately, a Campeau shareholder filed suit, charging Campeau...</i> (WSJ)
bill	24%	36%	<i>Currently the government charges nothing for such filings.</i> (WSJ)
credit card	0%	2%	<i>Many auto dealers now let buyers charge ... their purchase on the American Express card...</i> (WSJ)
TOTAL	100%	100%	

• **Verb senses have different subcategorization probabilities.**

Senses of charge	that-S	NP	NP PP	passive	Other
appoint	0%	0%	0%	4%	0%
accuse	18%	0%	12% (with)	24%	2%
bill	0%	9%	24% (for)	1%	1%
credit card	0%	0%	2% (on)	0%	0%

• **Controlling for verb sense improves agreement.**

Verb	Cosine (all senses combined)	Cosine (most common sense)
pass	0.75	0.95
charge	0.65	0.80
jump	0.50	0.59

Experimental Biases

• **Animate subject** (e.g. Garmsey) can bias the use of the verb, subcategorization - worry switches preference:

→ *Samantha worried that trouble was coming in waves.* (Garmsey - SC preference)

→ *Her words remained with him, worrying him for hours.* (BC - NP preference)

• Experimentally provided **setting/topic** (e.g. CFJCF) can bias the sense of the verb, subcategorization:

→ BC uses *pass* to mean *pass a test* about 1% of the time, → CFJCF data *pass* means *pass a test* 55% of the time, when the provided setting is school.

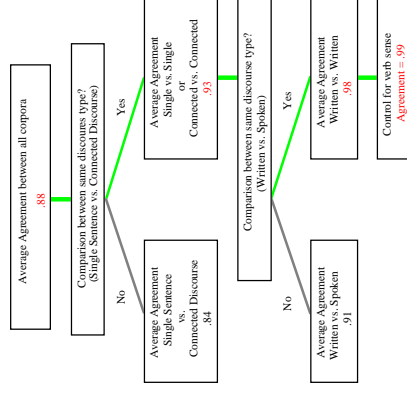
• Experimentally provided **verb tense** (e.g. Garmsey) can bias the aspectual reading of the verb, subcategorization:

→ *Debbie guessed the answer.* (Garmsey, perfective, DO) → *Poiter, guesses the poor man's mother died.* (BNC, imperfective, SC)

Combined Influence

- Lemma Argument Probability
- + Contextual Influence
- Observed Subcat Probability

Evidence that a single verb sense has a single subcategorization vector: The agreement between subcategorization vectors iteratively improves as we control for more factors. This data is for *learn*, which is one of the few verbs that appears in all five corpora.



Conclusion

- Lemma Argument Probability Hypothesis: Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.
- Probabilistic Combination Hypothesis: Observed core probabilities and contextual influence.
- Psychological models and studies must take word sense and context into account.

Acknowledgments

- Thanks to Charles Clifton for providing the original data from Comminne et al. (1984).
- Thanks to Sue Garmsey for providing the original data from Garmsey et al. (1997).
- This project was supported by the generosity of the NSF via NSF IIS-9733067, NSF IRI-9704046, NSF IRI-9618838 and the Committee on Research and Creative Work at the graduate school of the University of Colorado, Boulder. Many thanks to Giulia Benigni, Charles Fillmore, Susanne Gahl, Adele Goldberg, Michelle Gregory, Uli Heid, Paola Merlo, Neal Perlmutter, Bill Raymond, Philip Resnik, and two anonymous reviewers.