Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: The Role of Verb Sense

[1]Douglas Roland, [1,2,3]Daniel Jurafsky, [1,3]Lise Menn, [4]Susanne Gahl, [1]Elizabeth Elder and [1]Chris Riddoch

[1]Department of Linguistics, [2]Department of Computer Science, [3]Institute of Cognitive Science
University of Colorado
Boulder, CO 80309-0295
{douglas.roland, jurafsky, lise.menn, elizabeth.elder, christopher.b.riddoch}@colorado.edu

[4]Department of Linguistics
Harvard University
Cambridge MA 02138
sgahl@fas.harvard.edu

Abstract

We explore the differences in verb subcategorization frequencies across several corpora in an effort to obtain stable cross corpus subcategorization probabilities for use in norming psychological experiments. For the 64 single sense verbs we looked at, subcategorization preferences were remarkably stable between British and American corpora, and between balanced corpora and financial news corpora. Of the verbs that did show differences, these differences were generally found between the balanced corpora and the financial news data. We show that all or nearly all of these shifts in subcategorization are realised via (often subtle) word sense differences. This is an interesting observation in itself, and also suggests that stable cross corpus subcategorization frequencies may be found when verb sense is adequately controlled.

Introduction

Verb subcategorization probabilities play an important role in both computational linguistic applications (e.g. Carroll, Minnen, and Briscoe 1998, Charniak 1997, Collins 1996/1997, Joshi and Srinivas 1994, Kim, Srinivas, and Trueswell 1997, Stolcke et al. 1997) and psycholinguistic models of language processing (e.g. Boland 1997, Clifton et al. 1984, Ferreira & McClure 1997, Fodor 1978, Garnsey et al. 1997, Jurafsky 1996, MacDonald 1994, Mitchell & Holmes 1985, Tanenhaus et al. 1990, Trueswell et al. 1993).

Previous research, however, has shown that subcategorization probabilities vary widely in different corpora. Studies such as Merlo (1994), Gibson et al. (1996), and Roland & Jurafsky (1997) have found subcategorization frequency differences between traditional corpus data and data from psychological experiments. Biber (1993) and Biber et al. (1998) have shown that that word frequency, word sense (as defined by collocates), the distribution of synonymous words and the use of syntactic structures varies with corpus genre. Roland & Jurafsky (1998, 2000 in press) showed that there were subcategorization frequency differences between various written and spoken corpora, and furthermore showed that that these subcategorization frequency differences are caused by variation in word sense as well as genre and discourse type differences among the corpora.

While the subcategorization probabilities in a computational language model can be adjusted to match a particular corpus, cross corpus differences in such probabilities pose an important problem when using corpora for norming psychological experiments. If each corpus generates a separate set of probabilities, which probabilities are the correct ones to use as a model of human language processing?

In an attempt to use corpora to provide norming data for 64 verbs for experimental purposes, we investigate in detail how verb frequencies and verb subcategorization frequencies differ among three corpora: the British National Corpus

(BNC), the Wall Street Journal corpus (WSJ), and the Brown Corpus (Brown). For the 64 verbs, we randomly selected a set of sentences from each corpus and hand-coded them for transitivity, passive versus active voice, and whether the selected usage was an instance of the most common sense of the verb.

We then ask two questions: Do these verbs have the same subcategorization probabilities across corpora, and, when there are differences, what is the cause. If a set of factors causing the differences can be identified and controlled for, then a stable set of cross-corpus probabilities suitable for norming psychological experiments can be generated.

While previous work has shown that differences between corpora do exist, and that word sense differences play a large role in realising these differences, much less is known about the effect of other factors on subcategorization variation across corpora. For example, are there gross subcategorization differences between British and American English? To what extent does the business-genre nature of the Wall Street Journal corpus affect subcategorization probabilities? Finally, while Roland and Jurafsky (2000 in press) suggested that sense differences played a major role in subcategorization biases, they were only able to test their hypothesis on a small number of verbs.

Our eventual goal is an understanding of many levels of verb differences across corpora, including verb frequency, frequency of transitive versus intransitive uses, frequency of other subcategorization frames, and frequency of active versus passive use. This paper reports our preliminary results on the first two of these issues. Verb usage was surprisingly unaffected by differences between British and American English. Those differences that did occur seem mostly to be caused by differences in the distribution of verb senses across corpora. The business-genre nature of the Wall Street Journal corpus caused certain verbs to appear more often in particular senses that had a strong effect on its subcategorization frequencies. Even after controlling for the broad sense of the verb, we found subcategorization differences caused by the "micro-differences" in sense, including quite specific arguments to the verb.

1 Data

Data for 64 verbs (shown in Table 1) was collected from three corpora; The British National Corpus (BNC) (http://info.ox.ac.uk/bnc/index.html), the Penn Treebank parsed version of the Brown Corpus (Brown), and the Penn Treebank Wall Street Journal corpus (WSJ) (Marcus et al. 1993). The 64 verbs were chosen on the basis of the requirements of separate psychological experiments including having a single dominant sense, being easily imagable, and participating in one of several subcategorization alternations. A random sample of 100 examples of each verb was selected from each of the three corpora. When the corpus contained less than 100 tokens of the verb, as was frequently the case in the Brown and WSJ corpora, the entire available data was used. This data was coded for several properties: Transitive/Intransitive, Active/Passive, and whether the example involved the major sense of the verb or not. The BNC data was coded entirely by hand, while the Brown and WSJ was hand coded after a first pass of subcategorization labelling via a tgrep search string algorithm. The same coder labelled the data for all three corpora for any given verb, in order to reduce any problems in intercoder reliability.

adjust, advance, appoint, arrest, break, burst, carve, crack, crumble, dance, design, dissolve, distract, disturb, drop, elect, encourage, entertain, excite, fight, float, flood, fly, frighten, glide, grow, hang, harden, heat, hurry, impress, jump, kick, knit, lean, leap, lecture, locate, march, melt, merge, mutate, offend, play, pour, race, relax, rise, rotate, rush, sail, shut, soften, spill, stand, study, surrender, tempt, terrify, type, walk, wander, wash, watch

Table 1- 64 verbs chosen for analysis

2 Verb Frequency

Because word frequency is known to vary with corpus genre, we used the frequency differences for our target verbs as a measure of corpus difference. We would expect factors such as corpus genre (Business for WSJ vs. mixed for BNC and Brown), American vs. British English, and the era the corpus sample was taken in to influence word frequency.

We calculated the frequencies for each verb, and used Chi Square to test whether the difference in frequency was significant for each corpus pairing. We then counted the number of verbs that showed a significant difference using $p = 0.05$ as a cut-off point. This result is shown in Table 2. Although there were verbs that had a significant difference in distribution between the two mixed genre corpora (BNC, Brown), there were more differences in word frequency between the general corpora and the business corpus. The difference between the BNC/Brown comparison and the BNC and Brown vs. WSJ comparison is significant (Chi Square, $p < .01$).

BNC vs Brown	*BNC vs WSJ*	*Brown vs WSJ*
30/64	46/64	46/64

Table 2 – Number of verbs showing a significant difference in frequency between corpora.

Table 3 shows the list of words that were significantly more frequent in both of the general corpora than they were in the business oriented corpus. Notice that most of the verbs describe leisure activities.

amuse, boil, burst, dance, disturb, entertain, frighten, hang, harden, hurry, impress, knit, lean, paint, play, race, sail, stand, tempt, walk, wander, wash, watch

Table 3 - Verbs which BNC and Brown both have more of than WSJ:

Alternatively, when one looks at the words that had a significantly higher frequency in the WSJ corpus than in either of the other corpora (Table 4), one finds predominately verbs that can describe stock price changes and business transactions.

adjust, advance, crumble, drop, elect, fall, grow, jump, merge, quote, rise, shrink, shut, slip

Table 4 - Verbs which WSJ has more of than both Brown and WSJ:

We are currently examining the nature of the differences between the British and American corpora.

3 Subcategorization Frequency

3.1 Methodology:

For the second experiment, we coded the examples of the 64 verbs from each of the three corpora for transitivity. We counted any use with a direct object as transitive, and any other use, such as with a prepositional phrase, as intransitive. Passive uses were also included in the transitive category. Examples (1) and (2) illustrate intransitive uses, example (3) illustrates transitive (and active) while examples (4) and (5) illustrate transitive (and passive) uses of the verb ‘race’.

(1) Pretax profits <u>dropped</u> by 37 million.
(2) Something <u>dropped</u> to the floor.
(3) Lift them from the elbows, and then <u>drop</u> them down to the floor.
(4) Plans for an OSF binary interface have been <u>dropped</u>.
(5) It was ... the tinsel paper <u>dropped</u> by bombers.

Roland and Jurafsky (2000 in press) showed that verb sense can affect verb subcategorization. We therefore controlled for verb sense by only including sentences from the majority sense of the verb in our counts. For example, we did not include instances of drop which were phrasal verbs with distinct senses like "drop in" or "drop off". We did however, include metaphorical extensions of the main sense, such as a company "dropping a product line". We thus used a broadly defined notion of sense rather than the more narrowly defined word senses used in some on-line word sense resources such as Wordnet. This was partly for logistic reasons, since such fine-grained senses are very hard to code, and partially because we suspected that very narrowly defined senses frequently have only one possible subcategorization. Coding for such senses would have thus biased our experiment strongly toward finding a strong link between sense and subcategorization-bias.

We calculated transitivity biases for each of the 64 verbs in each of the three corpora. We classed the verbs as high transitivity if more than 2/3 of the tokens of the major sense were transitive, low transitivity if more than 2/3 of the tokens of the major sense were intransitive, and as mixed otherwise. We removed from consideration any token of the verb which was not used in its major sense. If subcategorization biases are related to verb sense, we would expect the transitivity biases to be stable across corpora once secondary senses are removed from consideration.

3.2 Results:

Nine of the 64 verbs, shown in Table 5, had a significant shift in transitivity bias. These verbs had a different high/mixed/low transitivity bias in at least one of the three corpora.

Verb	BNC transitivity	Brown transitivity	WSJ transitivity
advance	mixed (48%)	mixed (65%)	low (19%)
crack	mixed (58%)	mixed (58%)	high (86%)
fight	low (29%)	mixed (49%)	high (64%)
float	low (22%)	low (11%)	mixed (44%)
flood	mixed (52%)	high (100%)	high (100%)
relax	low (27%)	low (30%)	mixed (65%)
soften	high (71%)	high (70%)	mixed (43%)
study	high (84%)	mixed (39%)	high (92%)
surrender	mixed (48%)	mixed (39%)	high (73%)

Table 5 – Transitivity bias in each corpus

3.3 Discussion:

In general, these shifts in transitivity were a result of the verbs having differences in sense between the corpora such that the senses had different subcategorizations, but were still within our broadly defined 'main sense' for that verb.

For seven out of the nine verbs, the shifts in transitivity are a result of differences between the WSJ data and the other data, which are a result of the WSJ being biased towards business-specific uses of these verbs. For example, in the BNC and Brown data, 'advance' is a mixture of transitive and intransitive uses, shown in (6) and (7), while intransitive share price changes (8) dominated in the WSJ data.

(6) BNC intransitive: In films, they advance in droves of armour across open fields …
(7) BNC transitive: We have advanced "moral careers" as another useful concept ...
(8) WSJ intransitive: Of the 4,345 stocks that T changed hands, 1,174 declined and 1,040 advanced.

'Crack' is used to mean 'make a sound' (9) or 'break' (10) in the Brown and BNC data (both of which have transitive and intransitive uses), while it is more likely to be used to mean 'enter or dominate a group/market' (transitive use) in the WSJ data; (11) and (12).

(9) Brown intransitive: A carbine cracked more loudly …
(10) Brown intransitive: Use well-wedged clay, free of air bubbles and pliable enough to bend without cracking.
(11) WSJ transitive: But the outsiders haven't yet been able to crack Saatchi's clubby inner circle, or to have significant influence on company strategy.
(12) WSJ transitive: … big investments in "domestic" industries such as beer will make it even tougher for foreign competitors to crack the Japanese market.

'Float' is generally used as an intransitive verb (13), but must be used transitively when used in a financial sense (14).

(13) Brown intransitive: The ball floated downstream.
(14) WSJ transitive: B.A.T aims to … float its big paper and British retailing businesses via share issues to existing holders.

'Relax' is generally used intransitively (15), but is used transitively in the WSJ data when discussing the relaxation of rules and credit (16).

(15) BNC intransitive: The moment Joseph stepped out onto the terrace the worried faces of Tran Van Hieu and his wife relaxed with relief.
(16) WSJ transitive: Ford is willing to bid for 100% of Jaguar 's shares if both the government and Jaguar shareholders agree to relax the anti-takeover barrier prematurely.

'Soften' is generally used transitively (17), but is used intransitively in the WSJ data when discussing the softening of prices (18) and (19).

(17) Brown transitive: Hardy would not allow sentiment to soften his sense of the irredeemable pastness of the past, and the eternal deadness of the dead.
(18) WSJ intransitive: A spokesman for Scott says that assuming the price of pulp continues to soften, "We should do well."
(19) WSJ intransitive: The stock has since softened, trading around $25 a share last week and closing yesterday at $23.00 in national over-the-counter trading.

'Surrender' is used both transitively (20) and intransitively (21), but must be used transitively when discussing the surrender of particular items such as 'stocks' (22) and (23).

(20) BNC transitive: In 1475 Stanley surrendered his share to the crown…
(21) Brown intransitive: … the defenders, to save bloodshed , surrendered under the promise that they would be treated as neighbors
(22) WSJ transitive: Holders can … surrender their shares at the per-share price of $1,000, plus accumulated dividends of $6.71 a share.
(23) WSJ transitive: … Nelson Peltz and Peter W. May surrendered warrants and preferred stock in exchange for a larger stake in Avery 's common shares.

The verb 'fight' is the only verb that has a different transitivity bias in each of the three corpora; with all other verbs, at least two corpora share the same bias. In the WSJ, fight tends to be used transitively, describing action against a specific entity or concept (24). In the other two corpora, there are more descriptions of actions for or against more abstract concepts (25) and (26). In addition, the WSJ differences may further be influenced by a journalistic style practice of dropping the preposition 'against' in the phrase 'fight against'.

(24) WSJ transitive: Los Angeles County Supervisor Kenneth Hahn yesterday vowed to fight the introduction of double-decking in the area.
(25) BNC intransitive: He fought against the United Nations troops in the attempted Katangese secession of nineteen sixty to sixty-two.
(26) Brown intransitive: But he would fight for his own liberty rather than for any abstract principle connected with it -- such as "cause".

The verb 'study' is generally transitive (27), except in the Brown data, where study is frequently used with a prepositional phrase (28) or to generically describe the act of studying (29). We are currently investigating what might be causing this difference; possible candidates include language change (since Brown is much older than BNC and WSJ), British-American differences, or micro-sense differences.

(27) BNC transitive: A much more useful and realistic approach is to study recordings of different speakers' natural, spontaneous …
(28) Brown intransitive: In addition, Dr. Clark has studied at Rhode Island State College and Massachusetts Institute of Technology.
(29) Brown intransitive: She discussed in her letters to Winslow some of the questions that came to her as she studied alone.

The verb 'flood' is used intransitively more often in the BNC than in the other corpora. The Brown and WSJ uses tend to be transitive non-weather uses of the verb flood (30) and

(31), while the BNC uses include more weather uses, which are more likely to be intransitive (32). We are investigating whether this is a result of the BNC discussing weather more often, or a result of which particular grammatical structures are used to describe the weather floods in British and American English.

(30) WSJ transitive: Lawsuits over the harm caused by DES have flooded federal and state courts in the past decade.
(31) Brown transitive: The terrible vision of the ghetto streets flooded his mind.
(32) BNC intransitive: … should the river flood, as he 'd observed it did after heavy rain, the house was safe upon its hill.

Conclusion

The goal of the work performed in this paper was to find a stable set of transitivity biases for 64 verbs to provide norming data for psychological experiments.

The first result is that 55 out of 64 single sense verbs analyzed did not change in transitivity bias across corpora. This suggests that for our goal of providing transitivity biases for single sense verbs, the influence of American vs. British English and broad based vs. narrow corpora may not be large. We would, however, expect larger cross corpus differences for verbs that are more polysemous than our particular set of verbs.

The second result is that for the 9 out of 64 verbs that did change in transitivity bias, the shift in transitivity bias was largely a result of subtle shifts in verb sense between the genres present in each corpus. These two results suggest that when verb sense is adequately controlled for, verbs have stable subcategorization probabilities across corpora.

One possible future application of our work is that it might be possible to use verb frequencies and subcategorization probabilities of multi-sense verbs can be used to measure the degree of difference between corpora.

Acknowledgements

This project was partially supported by NSF BCS-9818827 and IRI-9618838. Many thanks to the three anonymous reviewers.

References

Biber, D. (1988) *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D. (1993) *Using Register-Diversified Corpora for General Language Studies*. Computational Linguistics, 19(2), 219-241.

Biber, D, Conrad, S., & Reppen, R. (1998) *Corpus Linguistics*. Cambridge University Press, Cambridge.

Boland, J. (1997). *Resolving syntactic category ambiguities in discourse context: probabilistic and discourse constraints*. Journal of Memory and Language 36, 588-615.

Carrol, J., Minnen, G., & Briscoe, T. (1998). Can subcategorization probabilities help a statistical parser? In Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora, Montreal, Canada.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *AAAI-97*, Menlo Park. AAAI Press.

Clifton, C., Frazier, L., & Connine, C. (1984) *Lexical expectations in sentence comprehension*. Journal of Verbal Learning and Verbal Behavior, 23, 696-708.

Collins, M. J. (1996) *A new statistical parser based on bigram lexical dependencies*. In Proceedings of ACL-96, 184--191, Santa Cruz, CA.

Collins, M. J. (1997) *Three generative, lexicalised models for statistical parsing*. In Proceedings of ACL-97.

Ferreira, F., and McClure, K.K. (1997). *Parsing of Garden-path Sentences with Reciprocal Verbs*. Language and Cognitive Processes 12, 273-306.

Fodor, J. (1978). *Parsing strategies and constraints on transformations*. Linguistic Inquiry, 9, 427-473.

Garnsey, S. M., Pearlmutter, N. J., Myers, E. & Lotocky, M. A. (1997). *The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences*. Journal of Memory and Language 37, 58-93.

Gibson, E., Schutze, C., & Salomon, A. (1996). *The relationship between the frequency and the processing complexity of linguistic structure*. Journal of Psycholinguistic Research 25(1), 59-92.

Joshi, A. & B. Srinivas. (1994) *Disambiguation of super parts of speech (or supertags): almost parsing*. Proceedings of COLING '94.

Jurafsky, D. (1996) *A probabilistic model of lexical and syntactic access and disambiguation*. Cognitive Science, 20, 137-194.

Kim A, Srinivas B and Trueswell J (1997). *Incremental Processing Using Lexicalized Tree-Adjoining Grammar: Symbolic and Connectionist Approaches,* Conference on Computational Psycholinguistics, Berkeley, California, August 1997.

MacDonald, M. C. (1994) *Probabilistic constraints and syntactic ambiguity resolution*. Language and Cognitive Processes 9, 157-201.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A.. (1993) *Building a Large Annotated Corpus of English: The Penn Treebank*. Computational Linguistics 19.2:313-330.

Merlo, P. (1994). *A Corpus-Based Analysis of Verb Continuation Frequencies for Syntactic Processing*. Journal of Pyscholinguistic Research 23.6:435-457.

Mitchell, D. C. and V. M. Holmes. (1985) *The role of specific information about the verb in parsing sentences with local structural ambiguity.* Journal of Memory and Language 24, 542--559.

Roland, Douglas and Daniel Jurafsky. (2000 in press). Verb sense and verb subcategorization probabilities. In Paola Merlo and Suzanne Stevenson (Eds.) John Benjamins.

Roland, Douglas and Daniel Jurafsky. (1998). How verb subcategorization frequencies are affected by corpus choice. Proceedings of COLING-ACL 1998. p 1117-1121.

Roland, D. and Jurafsky, D. (1997) *Computing verbal valence frequencies: corpora versus norming studies*. Poster session presented at the CUNY sentence processing conference, Santa Monica, CA.

Stolcke, A., C. Chelba, D. Engle, V. Jimenez, L. Mangu, H. Printz, E. Ristad, R. Rosenfeld, D. Wu, F. Jelinek and S. Khudanpur. (1997*) Dependency Language Modeling.* Center for Language and Speech Processing Research Note No. 24. Johns Hopkins University, Baltimore.

Tanenhaus, M. K., Garnsey, S. M., & Boland, J. (1990). *Combinatory lexical information and language comprehension.* In Altmann, Gerry T. M. (Ed); et al; Cognitive models of speech processing: Psycholinguistic and computational perspectives. Cambridge, MA, USA: Mit Press.

Trueswell, J., M. Tanenhaus and C. Kello. (1993) *Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference from Garden-Paths.* Journal of Experimental Psychology: Learning, Memory and Cognition 19.3, 528-553