Reclassifying subcategorization frames for experimental analysis and stimulus generation

Paula Buttery[1], Andrew Caines[2]

[1] University of Cambridge, 9 West Road, Cambridge CB3 9DP, U.K.
[2] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, U.K.
paula.buttery@cl.cam.ac.uk, acaines@ebi.ac.uk

Abstract

Researchers in the fields of psycholinguistics and neurolinguistics increasingly test their experimental hypotheses against probabilistic models of language. VALEX (Korhonen, Krymolowski & Briscoe, 2006) is a large-scale verb lexicon that specifies verb usage as probability distributions over a set of 163 verb SUBCATEGORIZATION FRAMES (SCFs). VALEX has proved to be a popular computational linguistic resource and may also be used by psycho- and neurolinguists for experimental analysis and stimulus generation. However, a probabilistic model based upon a set of 163 SCFs often proves too fine grained for experimenters in these fields. Our goal is to simplify the classification by grouping the frames into genera---explainable clusters that may be used as experimental parameters. We adopted two methods for re-classification. One was a manual, linguistic approach derived from verb argumentation and clause features; the other was an automatic, computational approach driven from the graphical representation of SCFs for use in Natural Language Processing technology. The premise was not only to compare the results of two quite different methods for our own interest, but also to enable other researchers to choose whichever re-classification better suited their purpose (one being grounded purely in theoretical linguistics and the other in practical language engineering). The various classifications are available as a free online resource to researchers.

Keywords: Verb classification, Experiment design, Cognitive modelling

1. Introduction

1.1. Overview

We describe a dual-approach re-classification of the widely-used verb SUBCATEGORIZATION FRAME (SCF) types which were defined by Briscoe and Carroll (Briscoe & Carroll, 1997; Briscoe, 2000). This is a set of 163 frames which describe verbs and their argument structures, ranging from the straightforward subject-verb intransitive (Frame 22, 'she reads') to more complex clauses involving extraposition, clausal complements, etc.

A linguistic method for re-classification reduces the 163 SCFs into 12 clusters based on simple linguistic properties. These clusters are each considered a SCF genus and are presented in Table 1. A second, graphical method derives from Briscoe and Carroll's description of the SCFs as directed graphs of grammatical relations between words. Several taxonomies are produced over varying degrees of specificity of these graphs.

1.2. Motivation

Our input set of SCFs comes from Briscoe and Carroll's amalgamation of the frames which feature in the Alvey NL Tools (ANLT; Boguraev et al, 1987) and COMLEX Syntax dictionaries (Grisham, Macleod & Meyers, 1994). Briscoe and Carroll originally listed 160 SCFs (1997: 357). Briscoe later extended the list to 163 in an unpublished manuscript (2000).

In the field of computational linguistics this set of SCFs has been employed in the development of several natural language processing algorithms; for instance in the automatic identification of diathesis alternations (McCarthy, 2001). The SCFs have also been used to extend VerbNet with novel verb classes (Kipper et al., 2006) and to supplement Levin's verb class taxonomy (cf. Levin, 1993; Korhonen & Briscoe 2004). The lexical resource VALEX (a popular, automatically derived, large-scale verb lexicon) specifies verb usage as probability distributions over this set of 163 SCFs (Korhonen, Krymolowski & Briscoe, 2006).

Recently these SCFs have been integrated into cognitive models. For instance, in order to build more accurate models of first language acquisition, distributions of these SCFs have been to used describe changes in child language (Buttery, 2006). The SCFs have also been used to explain experimental findings within psycholinguistics (e.g. Devereux, Korhonen & Tyler, 2011) and neurolinguistics (e.g. Bozic et al., 2011).

The SCFs are extremely fine-grained, reflecting the many complexities and subtleties of verb argumentation. Some frames necessarily differ by as little as the presence or absence of a direct object (cf. Frame 123, 'it cost ten pounds', and Frame 124, 'it cost him ten pounds'), or the use of a particle in the verb phrase: Frame 125, 'it set him back ten pounds'.[1]

For researchers in psycholinguistics and neurolinguistics, experimental hypotheses are increasingly being tested against probabilistic models of language. However, a probabilistic model based upon a set of 163 SCFs often proves too fine grained for experimenters in these fields. For instance, when studying fMRI neuro-imaging data (where temporal quantisation is crude compared to the functional operation of the brain) researchers would rather use a small number of broader conceptual parameters in their models. Our goal is to simplify the classification by grouping the frames into

[1] A full list of SCFs are found in Briscoe's unpublished manuscript (Briscoe, 2000) or at http://www.wordiose.co.uk/resources.

genera---explainable clusters that may be used as experimental parameters.

We adopted two methods for re-classification. One was a manual linguistic approach derived from verb argumentation and clause features; the other was an automatic, computational approach derived from the graphical representations of the SCFs.[2] The premise was not only to compare the results of two quite different methods for our own interest, but also to enable other researchers to choose whichever re-classification better suited their purpose (one being grounded purely in theoretical linguistics and the other in practical language engineering).

2. Subcategorization frames

2.1. Argument vs. adjunct

Subcategorization frames relate to the demands made by verbs on the number and type of arguments they permit and require, including obligatory and optional arguments but **not** adjuncts. Arguments are understood to be selected by the verb and complete its meaning. Adjuncts are optional and serve to extend the meaning of the central predication. The argument-adjunct 'distinction' is in fact a continuum with considerable grey area in the middle. Somers (1984), for one, describes a six-point adjunct-argument scale:

i, integral arguments
Jon doesn't have a hope

ii, obligatory arguments
He beat me

iii, optional arguments
William drank lager

iv, middles
Vic worked in the garden

v, adjuncts
Harvey sat licking his paws

vi, extra-peripherals
Bobby can eat, as you know

Generally speaking, arguments of type i-iv are relevant to this work.

The set of 163 SCFs referred to in this paper are a superset of the SCFs found in the Alvey NL Tools (ANLT; Boguraev et al, 1987) and COMLEX Syntax dictionaries (Grisham, Macleod & Meyers, 1994). The COMLEX lexicographers distinguish adjuncts from arguments using a set of criteria and heuristics (Meyers, Macleod and Grisham 1994). For instance, they state that prepositional phrases headed by *to* tend to be arguments, whereas PPs expressing time, manner, or place are mostly adjuncts. They also state that adjuncts occur with a large variety of verbs at a similar frequency whereas arguments occur with a high frequency with specific verbs.

[2]These graphs, which represent the grammatical relations between lexical items within a SCF, are more generally utilised by Natural Language Processing technology

2.2. Argument valency

The examples in (1) demonstrate some of the ways that verbs vary in their number of obligatory arguments. In (1a), *surf* only selects one argument - the subject noun phrase. Here, *surf* is an intransitive verb and we assign it to Frame 22 INTRANS[3]. In (1b), *bought* selects not only a subject but also an object noun phrase *a juicer* to complete its meaning. Thus we assign *buy* to Frame 24 NP. In (1c), *put* requires a subject as well as an object noun phrase *Harvey* and prepositional phrase *on the floor*. We therefore recognize *put* as Frame 49, the NP-PP frame, in this context.

1a, Stephen surfs
Frame 22: INTRANS

1b, Andrew bought a juicer
Frame 24: NP

1c, Lindsay put Harvey on the floor
Frame 49: NP-PP

Verbs may be associated with more than one subcategorization frame, since they may vary in their argument requirements. For example, *surf* can also be transitive (2a), *buy* can be ditransitive (2b) and *put* may co-occur with a particle (2c).

2a, Stephen surfs the internet
Frame 24: NP

2b, Andrew bought me a juicer
Frame 37: NP-NP

2c, Lindsay put up with his foibles
Frame 76: PART-NP / NP-PART

Note that the SCFs abstract over specific lexically governed particles, prepositions and specific predicate selectional preferences.

3. Reclassifying the SCFs

3.1. Linguistic approach

To provide a descriptive but less fine grained classification, the SCFs were annotated for three features: (a) subject type, (b) valency, and (c) presence of clause-final verb phrase or clause. Each feature has three possible values, as set out in Table 2.

Code	Subject	Valency	Clause-final
0	Noun phrase	-	Zero
1	Verb phrase	One	Verb phrase
2	Clause	Two	Clause
3	-	Three	-

Table 2: Linguistic approach --- three criteria for composition of genera.

Subject type can occur as a noun phrase, verb phrase or clause. The frames are mono-, di- or tri-valent: *i.e.*, they

[3]The frame numbers are taken from the original Briscoe & Carroll classification, along with the COMLEX SCF name.

Subj	Args	Final	Description
0	1	0	intransitive
0	2	0	monotransitive
0	3	0	ditransitive
0	1	1	intransitive plus VP
0	2	1	monotransitive plus VP
0	1	2	intransitive plus clause
0	2	2	monotransitive plus clause
0	3	2	ditransitive plus clause
1	1	0	VP-subject intransitive
1	2	0	VP-subject monotransitive
2	2	0	clause-subject monotransitive
2	1	2	clause-subject intransitive plus clause

Table 1: Linguistic approach --- composition of genera required to account for the 163 SCFs.

feature one, two or three arguments of the verb, the subject being at least one of those. After these arguments, the frames can optionally close with a verb phrase or clause (the third value in this case being absence).

Any SCFs clustered with identical values were then grouped into a genus. It would have been feasible to produce twenty-seven genera (3^3) if all combinations of all values were employed. However, the reorganisation was a data-driven process. As it turned out, the input set of SCFs clustered into just twelve of the available genera, as shown in Table 1.

Although only twelve members of the paradigm are required here, twenty-seven genera are maximally available in the current model. This means that the set of SCFs and their genera are readily extendible in future, and moreover that the genera are adaptable to other languages which may require a different set of parameters (or spoken English which will at the very least require a zero value for valency; think of common phrases in speech such as *just coming*, *only joking*, etc). This will in turn increase the number of available genera and enhance their adaptability.

The descriptive labels for each genus draw on the notion of transitivity which in its abstract sense is a binary category indicating the presence (or not) of an object. But there is also an inherently countable aspect to the concept which is more relevant for our purposes. This relates to the traditional distinction which is made between verbs as intransitive (subject only; *i.e.*, mono-valent), mono-transitive (subject and object; *i.e.*, di-valent), and ditransitive (subject and two objects; *i.e.*, tri-valent). Thus, for this particular classification, we need only identify and count the arguments, rather than any attempt to distinguish between direct and indirect objects.

Grouping direct and indirect objects as one may seem unnecessarily crude at first, but there is no pressing need to make this distinction within the most general classification. The point of the exercise is in fact to abstract away sufficiently from the SCFs to produce a new and useful taxonomic level, and at the same time maintain necessary distinctions relating to valency, subject type and clause-final structure.

The full set of SCFs do indeed distinguish between direct and indirect objects: incorporating this information to the general classification results (via a fourth 'object type' feature) produces a classification with a significantly larger number of genera (where several SCFs exist as single-member genera).

The twelve genera specified here (as well as an extended set of genera that incorporate object type information) are listed with their associated SCFs in full at http://www.wordiose.co.uk/resources.

3.2. Computational approach

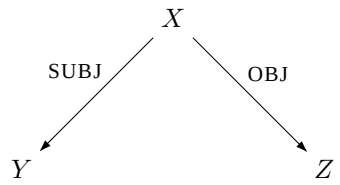
The computational re-classification of the SCFs is based on their definitions as expressed in the grammatical relations (GRs) notation used by the RASP system (Briscoe & Carroll, 2002; Briscoe et al., 2006). In a simplified form of this notation a transitive verb (Frame 24), exemplified by *Norman admires Alan*, may be defined as:

(SUBJ X Y)
(OBJ X Z)

which expresses that word Y is the subject of word X and word Z is the object of X. For our example sentence:

(SUBJ $admires$ $Norman$)
(OBJ $admires$ $Alan$)

As a graph we could think of this as a directed edge, labelled SUBJ, indicating a subject relationship from the word at vertex X to the word at vertex Y with a second edge, labelled OBJ, indicating an object relationship from vertex X to vertex Z:



We present several taxonomies based on such graphs, each with a differing level of specificity. The degree of specificity is provided by the amount of information available at a vertex or labelled edge.

Specificity is increased by:

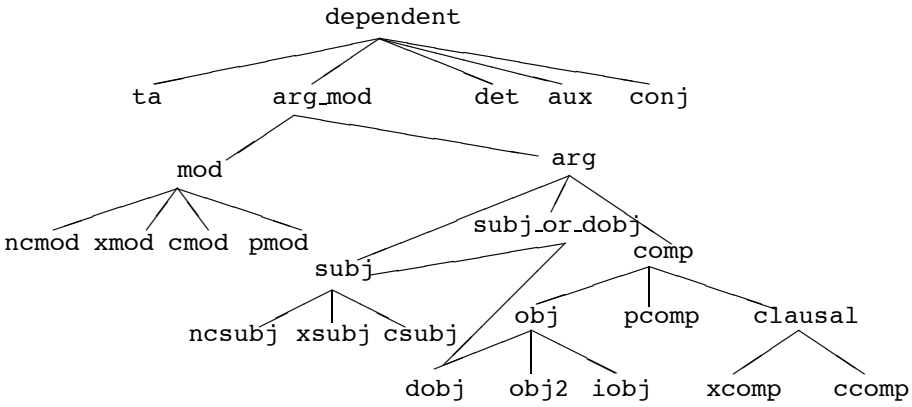


Figure 1: Briscoe and Carroll's grammatical relation hierarchy; employed as one method of varying specificity within the taxonomies

1. increasing the number of relationship types (and subtypes) that can be used in the definition of a SCF (i.e. increasing the number of possible labels for edges);

2. including lexical position information (i.e. connecting edges through vertices);

3. adding lexical type information at the vertices.

We are facilitated in deriving the specificity of the relationship types by the fact that Briscoe and Carroll's GRs have their own hierarchical structure (see Figure 1).
For an indication of how information specificity changes with taxonomy, we can track a particular SCF, Frame 55 or NP-TO-INF-VC, e.g. *they pushed her to leave*.

1 SUBJ OBJ CLAUSAL
2 SUBJ DOBJ CLAUSAL
3 (SUBJ X Y)
(OBJ X N)
(CLAUSAL X V)
4 (SUBJ X Y _)
(OBJ X N)
(CLAUSAL X V)

Figure 2: Example taxonomies with differing specificity for Frame 55

In the most general taxonomy (which is numbered 1 in Figure 2), Frame 55 is grouped within the broad category SUBJ OBJ CLAUSAL. This indicates that Frame 55 requires a subject, object and clausal complement. In the 2nd taxonomy the specificity has been increased by introducing subtypes of OBJ (i.e. DOBJ, IOBJ and OBJ2). It is now clear that the object required by Frame 55 is a DOBJ. In taxonomy 3 lexical information has been added. The variables X, Y, N and V specify both co-indexation and lexical type: X indicates the lexical head of the SCF; N indicates a required noun; V indicates a required verb; and Y indicates an unspecified lexical element. In taxonomy 4, finer grained grammatical relation details have been introduced: the _ in (SUBJ X Y _) indicates that the subject here is neither raised nor inverted and is not an underlying object (see (Briscoe, 2006) for more details).
Various combinations of specificity are of course possible. However, increasing the specificity has obvious consequences on the number of genera produced. For instance, Taxonomy 1 in Figure 2 has 8 genera, whereas taxonomy 4 has just fewer than 80 (about half as many as the fully specified set of 163 SCFs). When conducting experiments, researchers may select a taxonomy that best suits their needs in terms of information content and parameter space.
The taxonomies are available in full at http://www.wordiose.co.uk/resources.

4. Discussion

We have created several reclassifications of the 163 subcategorization frames by two methods. The first is based on linguistic principles and the second on specificity over graphical constructions used in natural language engineering. The manner of specificity of the latter has been informed through inter-disciplinary work with neuroscientists and psycholinguists, and the reclassifications are currently being used for experimental design and analysis. Examples include investigations into bilateral fronto-temporal systems and their role in speech comprehension (Bozic et al., 2011). In this work the syntactic complexity of verbs (defined through our taxonomies) is correlated with neural activity; the results suggest that the degree of dominance of lexical type (i.e. how confident a human subject is that a given stimulus word is a verb or a noun) interacts with syntactic complexity to determine a response in the fronto-temporal language network.
In other work (e.g. Devereux, Korhonen & Tyler, 2011; Devereux et al., 201) the taxonomies have been used to investigate parsing preferences for local ambiguities during sentence processing. The hypothesis here was that if the semantic class of the ambiguity-inducing noun had a stronger tendency to occupy a direct object position in a SCF than a subject position, then there would be a preference for gerundive readings. Whereas if the semantic class

of the noun had a stronger tendency to occupy the subject position there would be a preference for adjectival readings (e.g. *cooking pasta* versus *cooking husbands*). Frequencies for lexical items in the subject and direct object position were correlated over our taxonomies and used to interpret the results of sentence completion and sentence acceptability tasks. The results supported a lexicalist model of syntactic processing, where experience of verb lexico-syntactic behaviour (as revealed through corpus statistics) influences parsing preferences for local ambiguities during sentence processing.

Feedback from researchers in the fields of neurolinguistics and psycholinguistics is that they have found the flexible, computationally derived taxonomies (especially when combined with frequency distributions derived from corpora) to be a valuable tool for designing their stimuli and for manageably analysing results. Additionally, they have found the linguistically derived reclassification has facilitated their understanding of the genera in general and enabled more a more informed analysis.

The authors have made the resources discussed here together with supporting documentation available online at http://www.wordiose.co.uk/resources.

5. Acknowledgements

The first author was supported by Engineering and Physical Sciences Research Council (UK) grant number EP/F030061/1. Both authors thank the anonymous reviewers for their constructive criticism which was of much assistance in preparing this paper for submission. Finally, both authors sincerely acknowledge feedback, support and guidance from Mirjana Bozic, Ted Briscoe, Norman Cobley, Barry Devereux, Alan Horne, Jyothi Katuri, Barbara Lawn-Jones, Billi Randall, William Marslen-Wilson, Michael McCarthy, Peter Stoehr, Andrew Thwaites and Lorraine Tyler.

6. References

B.K. Boguraev, E.J. Briscoe, J.A. Carroll, D. Carter and C. Grover 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In: *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 193-200.

M. Bozic, E. Fonteneau, L.K. Tyler, B. Devereux, P. Buttery and W. Marslen-Wilson 2011. Grammatical categories in the fronto-temporal language network. In: *proceedings of 11th International Conference on Cognitive Neuroscience*, 120.

E.J. Briscoe 2000. Dictionary and System Subcategorisation Code Mappings. Unpublished manuscript, University of Cambridge Computer Laboratory.

E.J. Briscoe and J.A. Carroll 1997. Automatic extraction of subcategorization from corpora. In: *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, 356-363.

E.J. Briscoe and J.A. Carroll 2002. Robust accurate statistical annotation of general text. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, 1499-1504.

E.J. Briscoe, J.A. Carroll and R. Watson 2006. The Second Release of the RASP System. In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

E.J. Briscoe 2006. An introduction to tag sequence grammars and the RASP system parser. University of Cambridge, Computer Laboratory. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-662.pdf.

P.J. Buttery 2006. Computational models for first language acquisition. University of Cambridge, Computer Laboratory. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-675.pdf

B. Devereux, A. Korhonen and L.K. Tyler 2011. Parsing Sentences are Unlikely: Corpus-based Analyses of the Neural Processing of Verbs. In: *Proceedings of 11th International Conference on Cognitive Neuroscience*, 118.

B. Devereux, L.K. Tyler, P. Buttery and A. Korhonen 2011. The role of verb subcategorization frames and selectional preferences in sentence processing: an investigation using corpus-derived measures. Presented at: Multidisciplinary Workshop on the Mental Representation of Verbal Argument Structure.

R. Grishman, C. Macleod and A. Meyers 1994. COMLEX syntax: building a computational lexicon. In: *Proceedings of COLING*, 268-272. Association for Computational Linguistics.

K. Kipper, A.L. Korhonen, N. Ryant and M. Palmer 2006. Extending VerbNet with Novel Verb Classes. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

A.L. Korhonen 2002. Subcategorization acquisition. University of Cambridge, Computer Laboratory. http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-530.pdf

A.L. Korhonen and E.J. Briscoe 2004. Extended Lexical-Semantic Classification of English Verbs. In: *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*.

A.L. Korhonen, Y. Krymolowski and E.J. Briscoe 2006. A Large Subcategorization Lexicon for Natural Language Processing Applications. In: *Proceedings of the 5th international conference on Language Resources and Evaluation*.

B.C. Levin 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

D. McCarthy 2001. Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences. Ph.D thesis, University of Sussex.

A. Meyers, C. Macleod and R. Grishman 1994. Standardization of the complement adjunct distinction. In: *Proceedings of Euralex96*, 141-150.

H.L. Somers 1984. On the validity of the complement-adjunct distinction in valency grammar. *Linguistics* 22: 507-520.