"Category Families" for Categorial Grammars

Mary McGee Wood
Department of Computer Science
University of Manchester
Manchester M13 9PL U.K.
mary@cs.man.ac.uk

Abstract

Categorial Grammars (CGs; Wood 1993), grounded in algebra (Lambek 1958) and mathematical logic (Ajdukiewicz 1935), have rightly pushed to the limit the use of logically and algebraically justifiable rules for the combination and alternation of types in describing natural language. However, when TAG trees are mapped to CG categories, tree-families - linguistically well-motivated objects - can only be mapped to arbitrary category sets.

To capture predictable category alternations, such as noun / adjective alternations, or valency alternations for verbs, this paper proposes extending a CG with non-algebra-preserving rules, comparable to the "lexical redundancy rules" of early lexicalist theory. The theoretical argument is backed by an analysis of the degree of compaction which could be achieved by applying such rules to the CG "Large Lexicon" developed at IRCS, UPenn. The reduction which could be achieved both in the number of lexical entries and, more significantly, in the number of categories needed is considerable.

Redundancy rules in theory

CGs have always included both binary rules (such as function application and function composition) and unary (type-shifting) rules, and indeed the interactions between these two rule types have been involved in many debates within CG. The unary rules have been restricted to those which preserved algebraic identity: type-raising `NP` to `S/(S\NP)`, for example, does not in itself affect the descriptive power of the grammar. However, it is notorious that words can be highly ambiguous as to category, even in a phrase structure grammar with categories of a fairly coarse grain size (such as "verb"), but far more so in a CG. One of the central advances of the lexicalist movement in linguistic description (eg Bresnan (ed) 1982, Gazdar et al 1985) was the recognition and formalization of patterns in the lexicon such as active / passive alternation. Indeed it is ironic that the most extreme of lexicalist grammars has not adopted such lexical rules.

CGs could and, I believe, should have such type alternation rules. For example:

Nominals:
a lexical noun can also serve as a noun phrase, or a noun modifier or noun phrase modifier
`N` $\Longrightarrow$ `{NP, N/N, NP/NP}`

Passives:
a lexical verb will also have a passive form taking one fewer nominal complement
`(S\NP)/NP` $\Longrightarrow$ `S\NP`
`((S\NP)/NP)/NP` $\Longrightarrow$ `(S\NP)/NP`
etc.

Gerundives:
a verb (function into `S`) will also have a gerundive form (function into `NP`)
`(S\NP)/NP` $\Longrightarrow$ `(NP\NP)/NP`
`((S\NP)/NP)/NP` $\Longrightarrow$ `((NP\NP)/NP)/NP`
etc.

The exact semantics of the rewrite arrow is not at issue here. It is perhaps best taken as a well-formedness constraint or licensing statement along the lines of GPSG meta-rules: "if that is legal, so is this". Nor are we concerned with implementation details such as whether the rules cause expansion at run-time or compile-time. The claim is that these alternations are facts of natural language, and a linguistic theory must have rules to describe them, as indeed most linguistic theories do.

Redundancy rules in practice

The UPenn Combinatory CG "Large Lexicon" (Doran and Srinivas, forthcoming) was created by automatic translation from the large TAG lexicon developed by the TAG Group at the UPenn Institute for Research in Cognitive Science (XTAG Group 1995). TAG trees were mapped to CG categories, and the result modified by hand, principally by Christy Doran, B. Srinivas, and Mark Steedman. Some debugging remains to be done, so these figures are approximate:

36,950 entries
17,960 words
11 POS values
86 CG categories
120 CG category "families"
effectively about 110,000 entries (word / category pairs)

Category families are sets of categories which typically and predictably are assigned together to a word, causing the expansion from 37,000 word entries to 110,000 word / category pairs. In the original TAG lexicon, words are assigned *tree families*, which are linguistically well-motivated objects (Xia et al, in preparation). In the translation from TAG trees to CG categories, the motivation is lost, and we are left with seemingly arbitrary category sets. It is these which can be both motivated and compressed using redundancy rules.

Here are some example entries from the lexicon. (The index numbers serve to distinguish atoms within each complex category, and have no other significance. I give the corresponding TAG trees for the first entry only.)

Verbs: each verb stem has one or two block entries, with some redundancy in passive and gerundive categories:

```
INDEX:  crease/1
ENTRY:  crease
POS:    V
CAT:    S_0\NP_0
        NP_0\(NP_1/N_0)
        NP_0\NP_1

;;; Intransitives
GnxOV   NP_0\NP_1         #INTRANSger
InxOV   S_0\NP_0          #INTRANS
WOnxOV  S_0\NP_0          #INTRANS
nxOV    S_0\NP_0          #INTRANS
NOnxOV  S_0\NP_0          #INTRANS
DnxOV   NP_0\(NP_1/N_0)   #INTRANSger
                          #LagrpassNP_0
```

```
INDEX:  crease/2
ENTRY:  crease
POS:    V
CAT:    (S_0\NP_0)/NP_1
        (S_0\NP_0)/PP_0
        (NP_0\NP_1)/NP_2
        NP_0/NP_1
        N_0/N_1
FS:     #TRANS+
```

Nouns: each noun stem has four block entries, containing 12 categories (singular / plural x head noun / modifier, plus predicatives) which could be reduced to one:

```
INDEX:  Afghan/1
ENTRY:  Afghan
POS:    N
CAT:    (S_0\NP_0)\(NP_1/N_0)
        (S_0\S_1)\(NP_0/N_0)
FS:     #N_refl- #N_wh-

INDEX:  Afghan/2
ENTRY:  Afghan
POS:    N
CAT:    NP_0
        N_0
        N_0/N_1
        NP_0/NP_1
FS:     #N_refl- #N_wh-

INDEX:  Afghans/1
ENTRY:  Afghans
POS:    N
CAT:    (S_0\NP_0)\(NP_1/N_0)
        (S_0\S_1)\(NP_0/N_0)
FS:     #N_refl- #N_wh-

INDEX:  Afghans/2
ENTRY:  Afghans
POS:    N
CAT:    NP_0
        N_0
        N_0/N_1
        NP_0/NP_1
FS:     #N_refl- #N_wh-
```

Adjectives: each adjective has two block entries, containing four categories (singular / plural modifier, plus predicatives) which could be reduced to one:

```
INDEX:  Canadian/1
ENTRY:  Canadian
POS:    A
CAT:    NP_0/NP_1
        N_0/N_1
FS:     #A_WH-

INDEX:  Canadian/2
ENTRY:  Canadian
POS:    A
```

```
CAT:     S_0\NP_0
  (NP_0\NP_1)\((S_0\NP_2)/(S_1\NP_3))
FS:      #A_WH-
```

Since the exact figures for this sort of simple numerical compression are entirely dependent on incidental details of the composition of the original lexicon, it is more significant to look at the size of the set of categories used in the lexicon.

It is well known that CG categories are more detailed, and therefore more numerous, than the traditional categories of phrase structure grammars ("verb" becomes the set S\NP, (S\NP)/NP, ((S\NP)/NP)/NP, ..., etc.). It is less commonly observed that a single CG category can correspond to more than one PSG category, where different parts of speech have the same syntactic behaviour. For example,

S_0\NP_0

> Intransitive active
> *The scuffling and miaowing* **abated.**
> Transitive bare passive
> *The food* **was accepted.**
> Predicative adjective
> *That proposal is* **absurd.**
> Predicative nominal
> *Pepper is* **a tabby cat.**
> Predicative pp
> *The president is* **abroad.**

I refer to these as the *senses* of a category, and to a category with more than one sense as *ambiguous*. A *primary sense* is basic or irreducible, like the first sense (intransitive active) above. A *secondary sense* is a derived usage which could be predicted or derived by rule from some other category. Thus S_0\NP_0 (transitive bare passive) is derived from (S_0\NP_)/NP (transitive active) by a passive rule which systematically reduces the number of argument NPs to a verb by one. The three predicative senses are derived from basic adjectival, nominal, and prepositional categories by rules which are less neat schematically, but do make the appropriate predictions.

(Bear in mind that only the structural syntactic category itself is being considered here. Since TAG trees include part-of-speech information, "similar" looking trees are distinguished by the part-of-speech that anchors them. In CG categories, since part-of-speech information is not explicitly encoded, it appears that there are redundancies. However, as we saw above, lexical entries in the CCG Large Lexicon contain a POS field, so during lexical access, given a part-of-speech, there will not be any confusion of this nature.) Further, structurally identical categories will often be distinguished at a finer grain-size by having different features. The detailed form of any redundancy rules will have to include these.)

Although the proposed redundancy rules do give a worthwhile reduction in the number of categories needed, the number of senses which can be omitted, and the number of ambiguous categories, are more dramatically reduced.

The present CCG Large Lexicon category set includes:

> 86 categories, with
> 113 senses

of these, there are:

> 19 ambiguous categories, with
> 46 senses

By using redundancy rules to predict gerunds, passives, predicatives, and secondary nominal uses, we reduce this to

> 86 $\longrightarrow$ 65 categories, with
> 113 $\longrightarrow$ 73 senses

including:

> 19 $\longrightarrow$ 6 ambiguous categories, with
> 46 $\longrightarrow$ 14 senses.

The 40 senses eliminated (over one-third of the total) are made up of

> 12 gerunds
> 13 passives
> 13 predicatives
> 2 nominals

The 20 categories eliminated entirely include, for example:

((NP_0\NP_1)/NP_2)/NP_3

> Gerund of ditransitive
> *John* **giving** *the cats an unusually large breakfast kept them happy for a few hours.*

S_0/NP_0

> Predicative
> *Pepper is a tabby cat* - **What** *is Pepper?*

The thirteen ambiguous categories which become unambiguous include the example of S_0\NP_0 given above, which keeps only its primary sense of intransitive verb, losing four secondary senses, one passive and three predicative. When one considers that at present the first 15 words in the lexicon with this category are: *abate, abdicate, aberrant, abhorrent, abide, abject, able, abnormal, abominable, aboriginal, abort, abortive, above, abrasive, abroad* one advantage of the simplification is ovbvious. Similarly:

NP_0\NP_1
abate, abdicate, abide, abort, above, abroad, abscond, abstain, abut, accede, accelerate, accept, acclimatize, accord, accrue

Gerund of intransitive
the noise **abating**
Independent preposition
the stars **above**, *an Englishman* **abroad**

keeps only its prepositional sense and loses the gerundive.

The remaining ambiguities are entirely reasonable: for example,

(S_0\NP_0)/(S_1\NP_1)

Adverbs
Pepper **already** *was demanding breakfast.*
Auxiliary verbs
She **had** *prodded John's face several times.*

(S_0\NP_0)\(S_1\NP_1)

Adverbs
Pepper was demanding breakfast **already**.
Negation on auxiliaries
John did **not** *want to get up that early.*
Exhaustive PPs
He moved her **away**.

Redundancy rules will not only compress the explicitly given category set, but expand the set implicitly available. Crossing seven of the basic verb categories (intransitive, intrans + particle, intrans + adjective, transitive, trans + PP comp, trans + VP comp, trans + V comp) with five of the derived forms (active, bare passive, by-passive, gerund, gerund + determiner) should give 29 categories (as intransitives have no passive forms). Of these, only 18 are actually given in the current lexicon, presumably due to accidental gaps in the corpus data from which its parent TAG lexicon was originally derived.

Conclusion

This proposal will not be popular with the logical purists in the CG community. In language engineering terms, it will be necessary to control the appplicability of redundancy rules and to explore their effect on parsing. What I offer here is some quantified evidence, derived from a realistically large large lexicon intended for serious linguistic description, for the nature and scope of the benefits that a categorial grammar could gain from a systematic formalization of predictable lexical relations through lexical redundancy rules or category families.

Acknowledgements

This work rides on the shoulders of the people who developed the CCG Large Lexicon: I am deeply grateful to Christy Doran and B. Srinivas, in particular, for their generosity and support. Thanks also to David Brée, Jong C. Park, and Mark Steedman. Much of the credit is theirs; any errors or idiocies are mine.

References

Ajdukiewicz, K. 1935. "Die syntaktische Konnexität". *Studia Philosophica* 1:1-27; translated as "Syntactic connexion" in McCall (ed) *Polish Logic.* Oxford.

Bresnan, J. ed. 1982. *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge, Mass.

Doran, Christy and B. Srinivas. Forthcoming. "Developing a Wide-Coverage CCG System" To appear in a CSLI Volume of the *Proceedings of TAG+3*, ed. Anne Abeille and Owen Rambow.

Gazdar, G., E. Klein, G. Pullum and I. Sag. 1985. *Generalized Phrase Structure Grammar.* Blackwell, Oxford.

Lambek, J. 1958. "The Mathematics of Sentence Structure". *American Mathematical Monthly* 65:154-70; reprinted in Buszkowski, Marciszewski and van Benthem (eds) *Categorial Grammar.* John Benjamins, Amsterdam.

Wood, M.M. 1993. *Categorial Grammars.* Routledge, London.

Xia, Fei, Martha Palmer, K. Vijay-Shanker, Joseph Rosenzweig. In preparation. "Consistent Grammar Development Using Partial-Tree Descriptions for Lexicalized Tree-Adjoining Grammar".

The XTAG Group. 1995. "A Lexicalized Tree Adjoining Grammar for English". Technical Report IRCS 95-03, University of Pennsylvania. Updated version available at http://www.cis.upenn.edu/ xtag/tr/tech-report.h