Coordination Modules for a Crosslinguistic Grammar Resource

Scott Drellishak
University of Washington
sfd@u.washington.edu

Emily M. Bender
University of Washington
ebender@u.washington.edu

1 Background

The Grammar Matrix (Bender et al., 2002) is presented as an attempt to distill the wisdom of existing broad-coverage grammars and document it in a form that can be used as the basis for new grammars. The main goals of the project are: (i) to develop in detail semantic representations and in particular the syntax-semantics interface, consistent with other work in HPSG; (ii) to represent generalizations across linguistic objects and across languages; and (iii) to allow for very quick start-up as the Matrix is applied to new languages. The current Grammar Matrix release includes types defining the basic feature geometry and technical devices (e.g., for list manipulation), types associated with Minimal Recursion Semantics (see, e.g., (Copestake et al., 2003)), types for lexical and syntactic rules, and a hierarchy of lexical types for creating language-specific lexical entries, and links to the LKB grammar development environment (Copestake, 2002). It is, however, completely silent on the topic of coordination.

The next step in Matrix development is the creation of 'modules' to represent analyses of grammatical phenomena which differ from language to language, but nonetheless show recurring patterns. In this paper, we propose a design for a set of modules pertaining to coordination. Coordination is an especially important area to cover early on as coordinated phrases have a relatively high text frequency and thus could pose an important impediment to coverage in the development of Matrix-based grammars. In addition, while the world's languages evince a wide variety of coordination strategies, many of the challenges of providing grammatical analyses of coordination constructions are constant across all of the different strategies. Thus a relatively compact statement of the full set of possible modules is possible and the insights gained in existing work on coordination in the English Resource Grammar (version of 10/04, http://delph-in.net/erg; (Flickinger, 2000)) can be reasonably directly applied to other languages.

In this paper, we restrict our attention to *and* coordination but consider how coordination works for different phrase types as well as both 2-way and n-way coordination.[1] §2 provides a typological sketch of coordination strategies found in the world's languages. §3 motivates design decisions we have taken in this analysis. §4 presents a sample analysis of coordination in Ono. §5 discusses how we encode the information which can be compiled to create the types and instances needed for a particular grammar. Finally, in §6 we discuss further extensions to the grammatical analysis and issues of the user interface.

2 Typological Sketch

Across the world's languages, and across the phrase types within those languages, we find a wide variety of coordination strategies. These strategies can be classified along several dimensions; among these are the manner of marking, the location of the marking, and the etymological meaning of the mark.

The manner of marking coordination varies widely, and includes lexical, morphological, and phonological marking, as well as simple juxtaposition. The strategy most familiar from Indo-

[1] We leave for future work issues such as non-constituent coordination or the interaction of syncretism and coordination (e.g., (Beavers and Sag, 2004; Dalrymple and Kaplan, 2000)).

European languages is the use of a separate lexical item (e.g. English *and*). In some languages, coordination is not marked at all: the coordinands are merely juxtaposed. This occurs, for example, in the coordination of noun phrases in Abelam, a Sepik-Ramu language of Papua New Guinea:

(1) wʌny baləwʌny acʌ waryʌ.bər
that dog that pig fight
‘that dog and that pig fight’ (Laylock, 1965, 56)

Morphological marking generally involves inflecting one or more of the coordinands into some kind of conjunctive or continuative form. For example, in Kanuri (Nilo-Saharan) VPs can be coordinated by placing the earlier one in ‘conjunctive form’:

(2) kə̀rà̀zə̂ málə̀mrò wálwònò.
studied.CONJ malam became
‘He studied and became a malam.’
(Hutchison, 1981, 322)

In a few languages, coordination is marked by what appears to be a phonological alteration of the coordinands. For example, in Telugu (Dravidian), adjective phrases and noun phrases are coordinated by lengthening the final vowels in the coordinands:

(3) kamalaa wimalaa poDugu.
Kamala Vimala tall
‘Kamala and Vimala are tall.’
(Krishnamurti and Gwynn, 1985, 325)

Languages which require a special intonation contour to accompany coordination by juxtaposition are arguably using a phonological marking strategy as well. While ideally it would be very interesting to incorporate a model of prosody into grammar implementations, this is currently not feasible. Therefore, for present purposes, we will treat the juxtaposition strategy as though it had no overt marking.

Coordination strategies can also be classified by the location of the marking. In the simple case of two-way coordination, there are three positions where the marking may occur: before the first coordinand (initial), between the coordinands (medial), or after the second coordinand (final). In fact, the medial position is often more clearly associated with either the first or second coordinand, as a postfix or prefix respectively. In addition, languages vary in the number of marks used. If zero marks are used, we have the juxtaposition strategy, also referred to as *asyndeton*; if one mark is used, this is referred to as *monosyndeton*; if each coordinand is marked, this is referred to as *polysyndeton* (Haspelmath, 2000).

Finally, coordination strategies vary in the etymology of the marker. Some languages use an element related to the comitative marker and others an element not clearly related to anything else (Stassen, 2000). Rarer etymological sources include number words (Huánuco Quechua) and pronouns (Sedang).

Our intention with the coordination modules is to provide syntactic and semantic scaffolding powerful enough to deal with most or all of these structures, and flexible enough to be enhanced to cover other esoteric strategies that might be discovered.

3 Design Decisions

3.1 Category-specific Rules

It may seem desirable at first to have a single rule that covers the coordination of all phrase types. However, experience with detailed work on English (as represented by the English Resource Grammar) suggests that this is not practical, given our formalism and current assumptions about feature geometry. The core generalization[2] is that phrases of the same category can be coordinated to make a larger phrase of that category. Thus a common first-pass attempt at modeling coordination involves a rule that identifies HEAD and VAL values across the coordinands and the mother (see e.g., (Sag et al., 2003)). However, there are features which have been placed inside HEAD for independent reasons which need not be identified across coordinands, such as AUX:

(4) Kim slept and will keep on sleeping.

Further, there are differences in the semantic effects of coordination for individuals and events. In particular, nominal indices must be bound by quantifiers in MRS, leading NP and NOM coordination rules to introduce additional quantifiers. No such constraint holds for event indices.

Finally, there are idiosyncrasies to coordination in certain phrase types. A prime example here is the agreement features on coordinated NPs in English. For NPs coordinated with *and*, at least, the number

[2] This generalization is subject to several well known exceptions, which tend to have low text frequency.

of the conjoined phrase is always plural, and the person is the lesser of the person values of other coordinands (first person and second person give first person, etc.). In the context of our cross-linguistic analysis, we also find languages where the coordination strategy is different for different phrase types.

In light of these facts, the analysis is considerably simplified by positing separate rules for the coordination of different phrase types. These rules stipulate matching HEAD values, rather than identifying them. These rules are, of course, arranged into a hierarchy in which supertypes capture generalizations across all of the different coordination constructions.

3.2 Binary branching structure

Whether coordination involves binary branching or flat structure is a matter of much theoretical debate (see e.g., (Abeillé, 2003)). Rather than review those arguments here, we present two engineering considerations which support a binary branching analysis.

First, while the LKB allows rules with any given number of daughters, it does not permit rules with an underspecified number of daughters. This means that a rule like (5a) would have to be approximated via some number of rules with a specific arity (5b):

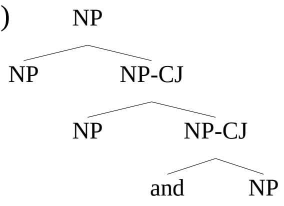
(5) a. NP → NP+ and NP

b. NP → NP and NP
NP → NP NP and NP
NP → NP NP NP and NP
. . .

With binary branching, in contrast, three rules produce an unlimited number of coordinands:

(6) NP-CJ → and NP (bottom coord rule)
NP-CJ → NP NP-CJ (mid coord rule)
NP → NP NP-CJ (top coord rule)

(7)



Second, there is the issue of 'promotion' of agreement features in coordinated NPs (and potentially other phrase types). In French, for example, the gender value of a coordinated NP is masculine iff at least one of the coordinands is. In order to state this constraint in this system, we'll need separate rule subtypes which posit [GEND *masc*] on the mother and on one daughter, leaving the other daughter unspecified.[3] In either system, this means doubling the number of rules, but the binary branching system starts out with fewer rules (and in fact, only the top and mid coordination rules need to be doubled, not the bottom coord rule). The flat structure system, on the other hand, potentially has a very large number of rules to start with. When we also consider promotion of person values, the number of rules involved gets larger, and the gain from the binary branching system becomes even clearer.

4 Sample Analysis

In this section, we provide a sketch of an analysis of coordination of verb phrases and noun phrases in Ono, a Trans-New Guinea language. As described by Phinnemore (1988), Ono verb phrases are coordinated by inflecting non-final verbs into a "medial" form, as in (8), while noun phrases are coordinated with the medial monosyndeton *so*, as in (9).

(8) mat-ine gelig-e taun-go ari
village-his leave-MED town-to go-MED
more zoma ka-ki so ea seu-ke
then sickness see-him-3sDS and there die-fp.-3s
'He left his village, went to town, and got sick and died there.' (Phinnemore, 1988, 109)

(9) koya so kezong-no numa len-gi
rain and clouds-ERG way block-3sDS
'Rain and clouds block the way...'
(Phinnemore, 1988, 100)

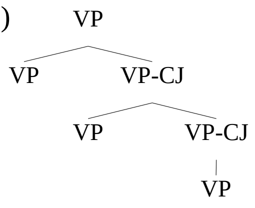
We handle these structures with six rules: *vp_top_coord_rule*, *vp_mid_coord_rule*, *vp_bottom_coord_rule*, *np_top_coord_rule*, *np_mid_coord_rule*, and *np_bottom_coord_rule*. The mid and top coord rules are non-headed rules with two daughters, one for each coordinand, called LCONJ-DTR and RCONJ-DTR. We assume additional boolean HEAD features COORD and (for verbs) MEDIAL. *vp_bottom_coord_rule* simply marks a [MEDIAL $-$] VP as coordinated (i.e. COORD $+$). The *vp_mid_coord_rule* will look something like the following:

[3] Dalrymple and Kaplan's (2000) set-based system for succinctly handling such facts is not currently available in the LKB.

$$\begin{bmatrix} \text{SYNSEM..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{verb} & \\ \text{VFORM} & \boxed{1} \\ \text{TAM} & \boxed{2} \\ \text{COORD} & + \end{bmatrix} \\ \text{VAL} & \boxed{3}\begin{bmatrix} \text{COMPS} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{LCONJ-DTR..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{verb} & \\ \text{VFORM} & \boxed{1} \\ \text{TAM} & \boxed{2} \\ \text{MEDIAL} & + \\ \text{COORD} & - \end{bmatrix} \\ \text{VAL} & \boxed{3} \end{bmatrix} \\ \text{RCONJ-DTR..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{verb} & \\ \text{VFORM} & \boxed{1} \\ \text{TAM} & \boxed{2} \\ \text{COORD} & + \end{bmatrix} \\ \text{VAL} & \end{bmatrix} \end{bmatrix}$$

This rule identifies several features of the coordinated VPs, marks the resulting phrase as coordinated, and takes a medial-form, noncoordinate left coordinand. This use of the COORD feature will enforce right-branching structure, so it is not necessary to specify MEDIAL on the mother node, which can only serve as the RCONJ-DTR of any further higher coordination. The Ono *vp_top_coord_rule* differs semantically from the mid rule in how it combines the semantic contributions of the coordinands, and differs syntactically from it only in that the mother node is [COORD −]. The structure assigned the coordination of three VPs, the first two of which are in medial form, is shown in (10), where VP-CJ is a VP marked [COORD +].

(10)



For noun phrases, we will need an additional lexical item *so* of HEAD type *conj*, and the *np_bottom_coord_rule* will combine *so* with an NP into a COORD-marked NP. The *np_mid_coord_rule* will look something like the following:

$$\begin{bmatrix} \text{SYNSEM..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{CASE} & \boxed{1} \\ \text{AGR} & \begin{bmatrix} \text{PER} & \boxed{2} \\ \text{NUM} & \textit{pl} \end{bmatrix} \\ \text{COORD} & + \end{bmatrix} \\ \text{VAL} & \boxed{3}\begin{bmatrix} \text{SPR} & \langle\,\rangle \\ \text{COMPS} & \langle\,\rangle \end{bmatrix} \end{bmatrix} \\ \text{LCONJ-DTR..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{CASE} & \boxed{1} \\ \text{AGR.PER} & \boxed{2} \\ \text{COORD} & - \end{bmatrix} \\ \text{VAL} & \boxed{3} \end{bmatrix} \\ \text{RCONJ-DTR..CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{noun} & \\ \text{CASE} & \boxed{1} \\ \text{AGR.PER} & \boxed{2} \\ \text{COORD} & + \end{bmatrix} \\ \text{VAL} & \boxed{3} \end{bmatrix} \end{bmatrix}$$

This rule identifies several features of the coordinated noun phrases, and constrains the mother to be plural, the mother and the RCONJ-DTR to be coordinated and the LCONJ-DTR to be not coordinated. The *np_top_coord_rule* will be similar, except that it combine the semantic contributions of all coordinands slightly differently, and will also mark the mother node [COORD −]. Based on these rules, the structure of a coordinated noun phrase made up of three NPs conjoined with a single *so* will look like (7) above, where NP-CJ is an NP marked [COORD +].

For languages with polysyndeton, the only modification to the rules in (6) is the omission of the mid rule, which results in the marking of coordination on each coordinand, because each additional NP will require one more bottom (and top) node:

(11)	NP-CJ	→ and NP	(bottom coord rule)
	NP	→ NP-CJ NP-CJ	(top coord rule)

5 Modularization

The intended goal of the coordination modules is to provide a basis for formal analyses for as wide a variety of languages as possible. However, we expect that we will be able to capture this variation based on a more limited set of semantic and syntactic rules. While it is not the case that all languages have the same number of or divisions between word classes, we expect to be able to capture the semantics of various phrase types in a language-independent way. The Matrix will provide coordination rules for phrases whose semantic contribution consists of individuals (e.g. noun phrases), events (e.g. verb phrases), modification of individuals (e.g. adjectives), modification of events (e.g. adverbs), and so forth.

In addition, we expect to find commonalities among the syntactic rules that can be factored out. For example, the parts of the VP and NP rules for Ono above that deal with the feature COORD can be adapted to deal with general asyndeton, monosyndeton, and polysyndeton coordination. All three strategies will have bottom and top coordination rules (with the mid rule only needed for monosyndeton), but the rules will vary slightly. The monosyndeton rules will look like the rules in (6) above; the polysyndeton rules will look like the rules in (11); and the asyndeton rules will look like (12).

(12) NP-CJ → NP (bottom coord rule)
NP → NP NP-CJ (top coord rule)

Different manners of marking coordination can be captured by varying the bottom rule. It can be either a rule that combines a separate lexical coordinator with the lowest coordinand, or else a non-branching rule triggered by a morphological feature.

Based on the answers to questions posed to the user about the facts of the language being analyzed, the semantic coordination rules and syntactic/morphological coordination rules will be cross-classified to produce a set of language-specific rules appropriate to the language at hand.

6 Conclusion and Outlook

We have presented an overview of an initial set of coordination modules for the Grammar Matrix. We believe that they are suited to providing syntactically and semantically valid analyses of the diverse coordination strategies in the world's languages. Furthermore, the factored representation given to the underlying types used to create language-specific coordination systems provides a means formalizing generalizations across languages.

The next steps for this project include: 1. Testing the coverage of the modules by deploying them in implemented grammars for a diverse range of languages. 2. Expanding the coverage to include other types of coordination (in the first instance, coordination with *or*, *but*, etc.). 3. Working out the user interface and in particular a set of questions and a protocol for presenting them to the linguist which covers the ground necessary to handle any given language while avoiding redundancy in any particular case.

References

Anne Abeillé. 2003. A lexicalist and construction-base approach to coordinations. In Stefan Müller, editor, *Proceedings of HPSG03*. CSLI, Stanford.

John Beavers and Ivan A. Sag. 2004. Ellipsis and apparent non-constituent coordination. In Stefan Müller, editor, *Proceedings of HPSG04*, pages 48–69. CSLI, Stanford.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix. *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*.

Ann Copestake, Daniel P. Flickinger, and Carl Pollard Ivan A. Sag. 2003. Minimal Recursion Semantics. An introduction.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI, Stanford.

Mary Dalrymple and Ronald M. Kaplan. 2000. Feature indeterminacy and feature resolution. *Language*, 76:759–798.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *NLE*, 6 (1):15 – 28.

Martin Haspelmath. 2000. Coordination. In Timothy Shopen, editor, *Language typology and linguistic description, 2nd edition*. Cambridge University Press, Cambridge.

John P. Hutchison. 1981. *A reference grammar of the Kanuri language*. University of Wisconsin - Madison, Madison, WI.

BH. Krishnamurti and J. P. L. Gwynn. 1985. *A grammar of modern Telugu*. Oxford University Press, Delhi.

D. C. Laylock. 1965. *The Ndu language family (Sepik district, New Guinea)*. Linguistic Circle of Canberra, Series C, No 1. The Australian National Library, Canberra.

Penny Phinnemore. 1988. Coordination in Ono. *Language and Linguistics in Melanesia*, 19:97–123.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Synactic Theory: A Formal Introduction*. CSLI, Stanford.

Leon Stassen. 2000. And-languages and with-languages. *Linguistic Typology*, 4:1–54.