Complex Case Phenomena in the Grammar Matrix

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

June 17, 2008

1 Introduction

The Grammar Matrix (Bender et al., 2002) is an attempt to provide a typologically-informed foundation for building grammars of natural languages in software. It includes a set of pre-defined types for lexical and syntactic rules, and a hierarchy of lexical types. It also provides a detailed syntax-semantics interface consistent with HPSG and Minimal Recursion Semantics (Copestake et al., 2005) and expressed in TDL (type description language) as interpreted by the LKB (Copestake, 2002). The primary purpose of the Matrix is to allow the rapid creation of new grammars based on the insights gained in the implementation of previous grammars.

The core of the Matrix is a set of types that are intended to be universal. Since there are linguistic phenomena that are widespread but not universal, the Matrix also includes "libraries" that consist of additional types covering non-universal phenomena (Bender and Flickinger 2005, Drellishak and Bender 2005). The Matrix also includes a customization system that prompts a linguist through a web-based questionnaire about a language, then creates a starter grammar, based on the Matrix and the appropriate libraries and tailored to the language. The current version of the questionnaire[1] includes mandatory sections on basic word order and basic lexical entries, and optional sections on sentential negation, coordination, and matrix yes/no questions.

This paper concerns efforts to add a library that supports case marking on verbal arguments. Development of such a library involves three steps. First, the typological range of case phenomena to be covered must be determined. Second, HPSG analyses must be developed for each of the possible case systems. Finally, these analyses must be "factored" into a set of consistent subanalyses that the customization system can put together in various ways to produce any grammar that can be described by answering the questionnaire. This paper will focus on the second step, the development of analyses, for several less-common case patterns, including split ergativity, Tagalog-style marking, and argument marking that is sensitive to scale hierarchies.

2 Case

Blake (2001) defines CASE as "a system of marking dependent nouns for the type of relationship they bear to their heads." This definition includes an extremely broad range of phenomena; in order to narrow the this range, the Grammar Matrix case library covers only case-marking of mandatory arguments of verbs. Even within this narrowed typological range, there exists considerable variation cross-linguistically.

Most notably, languages vary as to how intransitive and transitive clauses mark their arguments. Following Dixon (1994), I refer to the central grammatical roles of arguments as S (intransitive subject), A (transitive agent), and O (transitive patient or object). Some languages mark S and A with the same case, and O with another case; this is called the NOMINATIVE-ACCUSATIVE pattern. Other languages mark S and O the same, with A different; this is the ERGATIVE-ABSOLUTIVE pattern. Finally, some few languages mark all three roles differently; these are called TRIPARTITE languages.

All three types of NP argument marking can be handled on the verb lexical types using HPSG's ARG-ST feature (Manning and Sag, 1998) to constrain the argument structure, with the Argument Realization Principle providing the identities on the SUBJ and COMPS lists:

(1) Nominative-Accusative

verb-lex-item

$$\begin{bmatrix} \text{SYNSEM..HEAD.VAL.SUBJ} \left\langle \boxed{1}, \ldots \right\rangle \\ \text{ARG-ST} \left\langle \boxed{1} \left[\text{..HEAD.CASE} \quad \textit{nom} \right] \right\rangle \end{bmatrix}$$

intrans-verb-lex-item

$$\left[\text{SYNSEM..HEAD.VAL.COMPS} \left\langle \right\rangle \right]$$

trans-verb-lex-item

$$\begin{bmatrix} \text{SYNSEM..HEAD.VAL.COMPS} \left\langle \boxed{1} \right\rangle \\ \text{ARG-ST} \left\langle \left[\right], \boxed{1} \left[\text{..HEAD.CASE} \quad \textit{acc} \right] \right\rangle \end{bmatrix}$$

(2) Ergative-Absolutive

[1] http://www.delph-in.net/matrix/customize/matrix.cgi

intrans-verb-lex-item

$$\begin{bmatrix} \textsc{synsem..head.val.subj} \left\langle \boxed{1} \right\rangle \\ \textsc{arg-st} \left\langle \boxed{1} \begin{bmatrix} \textsc{..head.case} & abs \end{bmatrix} \right\rangle \end{bmatrix}$$

trans-verb-lex-item

$$\begin{bmatrix} \textsc{synsem..head.val} & \begin{bmatrix} \textsc{subj} & \left\langle \boxed{1} \right\rangle \\ \textsc{comps} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \textsc{arg-st} & \left\langle \begin{matrix} \boxed{1} \begin{bmatrix} \textsc{..head.case} & erg \end{bmatrix}, \\ \boxed{2} \begin{bmatrix} \textsc{..head.case} & abs \end{bmatrix} \end{matrix} \right\rangle \end{bmatrix}$$

(3) Tripartite

intrans-verb-lex-item

$$\begin{bmatrix} \textsc{synsem..head.val.subj} \left\langle \boxed{1} \right\rangle \\ \textsc{arg-st} \left\langle \boxed{1} \begin{bmatrix} \textsc{..head.case} & S \end{bmatrix} \right\rangle \end{bmatrix}$$

trans-verb-lex-item

$$\begin{bmatrix} \textsc{synsem..head.val} & \begin{bmatrix} \textsc{subj} & \left\langle \boxed{1} \right\rangle \\ \textsc{comps} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \textsc{arg-st} & \left\langle \begin{matrix} \boxed{1} \begin{bmatrix} \textsc{..head.case} & A \end{bmatrix}, \\ \boxed{2} \begin{bmatrix} \textsc{..head.case} & O \end{bmatrix} \end{matrix} \right\rangle \end{bmatrix}$$

The analysis of case in the Grammar Matrix case library also allows a variety of NP-marking strategies, including case-marking adpositions and morphological marking on nouns, determiners, or both. A discussion of these strategies is omitted here for lack of space.

2.1 Split Ergativity

Many languages are neither consistently ergative nor consistently accusative. Such languages are referred to as SPLIT ERGATIVE. In order to support them, the Matrix customization system must be able to create grammars in which more than one kind of marking, commonly the ergative and accusative patterns, co-exist.

Dixon (1994, 70) divides split ergative languages into four categories, based on how the split is conditioned:

1. Semantic nature of verb
2. Semantic nature of noun
3. Tense/aspect/mood of clause
4. Grammatical status of clause

The first type of split occurs in two subtypes. In one, called Split-S, the intransitive verbs are divided into two classes: those that take A-like marking on their single arguments and those that take O-like marking. I analyze this pattern as having one transitive verbs class with A- and O-marked argument, but two intransitive classes:

(4) *agent-intrans-verb-lex*

$$\begin{bmatrix} \textsc{arg-st} \left\langle \begin{bmatrix} \textsc{..head.case} & A \end{bmatrix} \right\rangle \end{bmatrix}$$

patient-intrans-verb-lex

$$\begin{bmatrix} \textsc{arg-st} \left\langle \begin{bmatrix} \textsc{..head.case} & O \end{bmatrix} \right\rangle \end{bmatrix}$$

The other subtype is called Fluid-S, in which the single argument of any intransitive verb can be marked like A or like O, depending on whether the subject controls the action or not: when a speaker marks an intransitive subject like A, this emphasizes the agency of the subject; when the subject is marked like O, this implies a lack of volition on the part of the subject. The semantic representation in grammars produced by the Matrix customization system do not presently have any way to show such a distinction; therefore, I analyze Fluid-S languages by simply specifying that the case of intransitive subjects is a supertype of A and O.

The second type of ergativity split is conditioned on the semantic nature of the nominal arguments. In such languages, certain kinds of NPs (e.g. pronouns) are marked in a nominative-accusative pattern while others (e.g. common nouns) are marked in an ergative-absolutive pattern. Furthermore, there exist languages of this type where the split is governed by a hierarchy, where what matters is the relative position of the NP arguments. §2.4 briefly describes my analysis of Fore, a language described as having a hierarchy-sensitive ergativity split.

The third type of split is conditioned on the tense, aspect, or mood of the verb. In many Iranian languages, for example, clauses in the past tense are marked in an ergative-absolutive pattern, while clauses in other tenses take nominative-accusative marking (Dixon, 1994, 100). The fourth type of split is conditioned on the grammatical status of the clause; that is, whether it is a main or subordinate clause.

I analyze the third and fourth types of split in the same way. The case type has (at least) four values: nominative, accusative, ergative, and absolutive. Verb lexical items have no case specified on their arguments; however, a set of mandatory lexical rules is used to constrain the CASE values on the ARG-ST list. For languages with the third type of split, it is the lexical rule that marks the conditioning feature (e.g. the past-tense morpheme) that constrains the CASE of the arguments. For languages with the fourth type of split, I make use of the Matrix's MC (main clause) feature, creating two non-spelling-changing lexical rules, one of which marks the clause as $\begin{bmatrix} \textsc{mc} + \end{bmatrix}$ with the appropriate case pattern, and the other as $\begin{bmatrix} \textsc{mc} - \end{bmatrix}$ with the other pattern.

2.2 Tagalog-type Languages

In some Austronesian languages, an interesting variant of verbal argument marking appears (Comrie, 1989, 120). In Tagalog (Austronesian, Philippines), a language of this type, every clause must have an NP argument marked with *ang*, which is referred to as a FO-

[2] It should be mentioned, however, that the term *focus* is here used rather differently than elsewhere in the linguistics literature.

[3] Comrie actually uses the terms *actor* and *undergoer*, but I use *agent* and *patient* here for consistency.

CUS marker (Comrie, 1989, 121).[2] In clauses with an agent and a patient[3], the other (non-*ang*-marked) NP is marked with *ng*. The verb in such clauses is marked by one of a set of affixes that tell how the *ang*- and *ng*-marked NPs should be interpreted, including agent-focus and patient-focus affixes. This pattern can be seen in the following examples:

(5) *Bumili* *ang* *babae* *ng* *baro*
bought-AGENT-FOC FOC woman PATIENT dress
'The woman bought a dress'

(6) *Bimili* *ng* *babae* *ang* *baro*
bought-PATIENT-FOC AGENT woman FOC dress
'A/the woman bought the dress' (Comrie, 1989, 121)

This manner of argument-marking is not straightforwardly accusative or ergative, instead constituting a distinct pattern. I analyze it as follows, using a slight modification of the analysis in §2. First, there exist adpositions that mark focus and non-focus. Next, every transitive verb lexical entry's ARG-ST is unspecified for case. For every type of focus-marking that can appear on a verb (including agent and patient focus), a lexical rule both applies the appropriate morphological marking and specifies the case of the arguments. The rules for agent- and patient-focus marking are:

(7) *agent-focus-verb-lex-rule*

$$\begin{bmatrix} \text{INPUT} & \left\langle \boxed{1}, \textit{tv-lex-item} \right\rangle \\ \text{OUTPUT} & \left\langle \begin{matrix} F_{af}(\boxed{1}), \\ \begin{bmatrix} \text{ARG-ST} \left\langle \begin{matrix} [\ldots\text{CASE} \;\; \textit{focus}], \\ [\ldots\text{CASE} \;\; \textit{non-focus}] \end{matrix} \right\rangle \end{bmatrix} \end{matrix} \right\rangle \end{bmatrix}$$

patient-focus-verb-lex-rule

$$\begin{bmatrix} \text{INPUT} & \left\langle \boxed{1}, \textit{tv-lex-item} \right\rangle \\ \text{OUTPUT} & \left\langle \begin{matrix} F_{pf}(\boxed{1}), \\ \begin{bmatrix} \text{ARG-ST} \left\langle \begin{matrix} [\ldots\text{CASE} \;\; \textit{non-focus}], \\ [\ldots\text{CASE} \;\; \textit{focus}] \end{matrix} \right\rangle \end{bmatrix} \end{matrix} \right\rangle \end{bmatrix}$$

2.3 Direct-inverse Languages

In languages with DIRECT-INVERSE marking, verbal arguments are marked in a pattern that is sensitive to a hierarchy. If the agent is ranked more highly than the patient, then the verb is in DIRECT form; if the patient is higher, the verb is in INVERSE form. For a concrete example, consider the Algonquian languages, where the hierarchy is primarily sensitive to person:

(8) 2nd > 1st > 3rd proximate > 3rd obviative

When a transitive clause contains two third-person arguments, one of them will be marked as proximate and the other as obviative to prevent ambiguity. The Algonquian proximate NP, according to (Dahlstrom, 1991, 91), is generally "the topic of the discourse" or "the focus of the speaker's empathy". The proximate NP is generally unmarked, while the obviative noun is marked by a suffix.

It is important to note that this is called a hierarchy, but it differs markedly from the sort of multiply-inheriting type hierarchies used in HPSG.[4] The hierarchy in (8) only establishes unstructured precedence relationships among the positions of the hierarchy; in contrast, HPSG-style type hierarchies establish tree-structured relationships among the items they contain. To avoid confusion, I will hereafter refer to hierarchies like (8) as SCALE HIERARCHIES.

The following examples from Fox (Algonquian) illustrate how argument marking works in a direct-inverse language:

(9) *ne* *-waapam-aa* *-wa*
1SG see-DIRECT 3
'I see him.'

(10) *ne* *-waapam-ek* *-wa*
1SG see-INVERSE 3
'He sees me.' (Comrie, 1989, 129)

Analyzing the direct-inverse pattern is challenging in the version of HPSG used in the Matrix (which, recall, is expressed in TDL and interpreted by the LKB system). For transitive verbs, it is necessary to have lexical rules for the direct and inverse forms that correctly constrain the verb's arguments. This could be expressed compactly if the formalism had some mechanism for stating scale-hierarchical constraints, something like:

(11) *direct-verb-lex-rule*

$$\begin{bmatrix} \text{INPUT} & \left\langle \boxed{1}, \ldots \right\rangle \\ \text{OUTPUT} & \left\langle F_{dv}(\boxed{1}, \begin{bmatrix} \text{ARG-ST} \left\langle \boxed{2}, \boxed{3} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix} \& \; \boxed{2} >> \boxed{3}$$

inverse-verb-lex-rule

$$\begin{bmatrix} \text{INPUT} & \left\langle \boxed{1}, \ldots \right\rangle \\ \text{OUTPUT} & \left\langle F_{iv}(\boxed{1}, \begin{bmatrix} \text{ARG-ST} \left\langle \boxed{2}, \boxed{3} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix} \& \; \boxed{2} << \boxed{3}$$

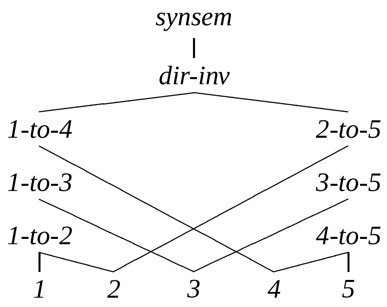
However, no such mechanism exists, so another method of analyzing scale hierarchies is required. It would obviously be possible to simply enumerate all possible combinations of pairs of positions on the scale hierarchy, creating a lexical rule for each pair, but this would mean having on the order of n^2 lexical rules for an n-position hierarchy. It would be better to somehow model the scale hierarchy with a type hierarchy.

Perhaps, noticing that it is necessary to address ranges of the scale hierarchy that start at the left or the

[4] This usage of *hierarchy*, it should be noted, has quite a long history in linguistics, and includes such well-known examples as the Noun Phrase Accessibility Hierarchy of Keenan and Comrie (1977)

right end, the scale hierarchy could be modeled using a type hierarchy like (12) (labeling the positions on the scale from 1 through 5), which is then used to constrain a series of lexical rules including those in (13) (which all derive from a single rule that applies the direct morphology to the verb):

(12)



(13) *direct-verb-lex-rule-1*
[ARG-ST ⟨*1*, *2-to-5*⟩]

direct-verb-lex-rule-2
[ARG-ST ⟨*1-to-2*, *3-to-5*⟩]

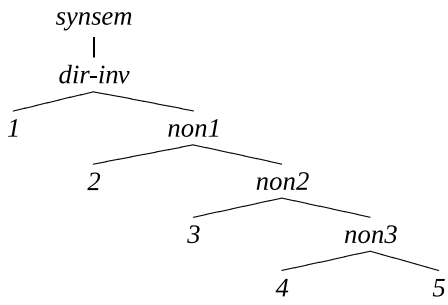
direct-verb-lex-rule-3
[ARG-ST ⟨*1-to-3*, *4-to-5*⟩]

direct-verb-lex-rule-4
[ARG-ST ⟨*1-to-4*, *5*⟩]

Unfortunately, this set of rules produces spurious ambiguity when applied to some sentences. While a sentence with, say, a subject from class 1 and an object from class 2 would be parsed only once, with *direct-verb-lex-rule-1* having applied to the verb, a sentence with a subject from class 1 and an object from class 5 would be parsed four times, once for each of the above rules.

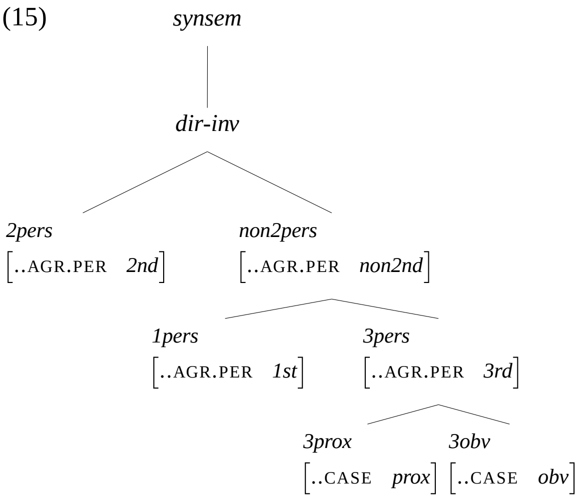
This problem can be addressed by revising the *dir-inv* type hierarchy. Instead of having ranges that extend from both ends, the revised hierarchy consists of pairs of types, one covering a single class in the scale hierarchy and the other the rest of the scale to the right, arranged into a right-branching tree:

(14)



In order to prevent spurious parses, the type hierarchy must constrain the appropriate syntactic features on both the leaves and the non-terminal nodes of the tree. For example, here are the type hierarchy (15) and lexical rules (16) for an Algonquian language with the scale hierarchy in (8):

(15)



(16) *direct-verb-lex-rule-1*
[ARG-ST ⟨*2pers*, *non2pers*⟩]

inverse-verb-lex-rule-1
[ARG-ST ⟨*non2pers*, *2pers*⟩]

direct-verb-lex-rule-2
[ARG-ST ⟨*1pers*, *3pers*⟩]

inverse-verb-lex-rule-2
[ARG-ST ⟨*3pers*, *1pers*⟩]

direct-verb-lex-rule-3
[ARG-ST ⟨*3prox*, *3obv*⟩]

inverse-verb-lex-rule-3
[ARG-ST ⟨*3obv*, *3prox*⟩]

Under this analysis, sentences will parse only once, solving the problem of spurious ambiguities. For example, a sentence with a verb in the direct form and a second-person agent will parse just once, regardless of the person and case of the patient, with *direct-verb-lex-rule-1* having applied to the verb. However, it is worth noting some drawbacks to this analysis. First, it is necessary to have, for a scale hierarchy with n positions, $2(n-1)$ lexical rules. Note also that the hierarchy in (15) is arbitrarily right-branching. A analysis could just as easily have been built around a left-branching hierarchy. Having two equally-valid analyses with nothing to choose between them may seem like luxury, but it could also be argued that it results from the inability of the formalism being used to compactly and efficiently express the linguistic generalization being analyzed.

2.4 Other Scale Hierarchies

Scale hierarchies affect the verbal argument marking patterns in other languages without direct-inverse marking on the verb. One example occurs in Fore (Trans-New Guinea), where the relative position of agent and patient on a scale hierarchy correlates with the presence or absence of a marker on the agent NP. The scale is:

(17) pron., name, kin term > human > anim. > inanim.

The operation of this hierarchy can be seen in the following examples:

(18) *yaga: wá* aegúye
pig man 3SG.hit.3SG
'The man kills the pig'

(19) *yaga:-wama wá* aegúye
pig-DLN man 3SG.hit.3SG
'The pig kills the man'

(20) *wa yága:-wama* aegúye
man pig-DLN 3SG.hit.3SG
'The pig kills the man' (Scott 1978, 116, Blake 2001, 122)

An extra suffix *wama* appears on the agent when it is lower on the hierarchy than the patient. Scott (1978) describes this pattern in a way that recalls a direct-inverse language with no marking on the verb, an overt proximate marker, and a zero-marked obviative. Blake (2001), on the other hand, describes the suffix as the ergative case marker, and therefore analyzes Fore as a language with split ergativity with the split conditioned on the hierarchy in (17).

Whether this implies an analysis as in §2.1 or in §2.3, it is necessary to translate the scale hierarchy in (17) into a type hierarchy. However, the pattern of features that distinguish the positions on the Fore scale (17) is quite different from that in the Algonquian scale (8). Rather than basing it on person and case, it is necessary in Fore to distinguish pronouns, names, and kin terms from common nouns, and to distinguish common nouns between humans, animates, and inanimates. Depending on whether and how the grammar is elsewhere sensitive to these distinctions, they could be modeled as a feature NTYPE of nominal heads, as a GENDER feature on nominal indices, or both.

3 Conclusion

In this paper I have described the analyses of a number of verbal argument marking patterns, including nominative-accusative, ergative-absolutive, tripartite, split ergative, Tagalog-type, and direct-inverse, that fall into the category of case. This development and implementation of such sets of analyses, where it must be possible to plug each analysis into a Matrix-based grammar, represents an instance of what could be called computational linguistic typology. That is, rather than analyzing languages separately, as syntacticians often do, or collecting descriptions of the range a phenomenon in the world's languages, as typologists do, in this project I aim to analyze in detail the typological range of a phenomenon (namely case) in a single framework, in the hope that such analysis will bring to light commonalities among human languages. This effort has already born some fruit. Notice, for example, that the analyses of several complex case patterns (e.g. split ergativity, Tagalog-type, and direct-inverse) can all be accomplished using single underlying verb lexical entries with a complex of mandatory lexical rules that produce the variation. Also notice that, in direct-inverse languages and in languages where ergativity splits are conditioned on a scale hierarchy, very similar HPSG type hierarchies of SYNSEMs can be used to model the behavior of the scale hierarchy. Hopefully, the implementation of other libraries for the Grammar Matrix and the resolution of any interactions that arise with existing libraries will reveal further generalizations.

References

Bender, Emily M. and Flickinger, Dan. 2005. Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05*, Jeju Island, Korea.

Bender, Emily M., Flickinger, Dan and Oepen, Stephan. 2002. The Grammar Matrix. In *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

Blake, Barry J. 2001. *Case, Second Edition*. Cambridge: Cambridge University Press.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology, Second Edition*. Chicago: University of Chicago Press.

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

Copestake, Ann, Flickinger, Dan, Pollard, Carl and Sag, Ivan A. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation* 3(2–3), 281–332.

Dahlstrom, Amy. 1991. *Plains Cree Morphosyntax*. New York: Garland Publishing.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Drellishak, Scott and Bender, Emily M. 2005. A Coordination Module for a Crosslinguistic Grammar Resource. In *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, Lisbon, Portugal.

Keenan, E. L. and Comrie, B. 1977. Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry* 8 (1), 63–99.

Manning, Christopher D. and Sag, Ivan A. 1998. Argument Structure, Valence, and Binding. *Nordic Journal of Linguistics* 21.

Scott, Graham. 1978. *The Fore Language of Papua New Guinea*. Canberra, Australia: Pacific Linguistics.