Complex case phenomena in the Grammar Matrix

Scott Drellishak
University of Washington

Proceedings of the 15th International Conference on
Head-Driven Phrase Structure Grammar

National Institute of Information and Communications Technology, Keihanna

Stefan Müller (Editor)

2008

Stanford, CA: CSLI Publications

pages 67–86

Drellishak, Scott. 2008. Complex case phenomena in the Grammar Matrix. In Stefan Müller (ed.), *Proceedings of the 15th International Conference on Head-Driven Phrase Structure Grammar, National Institute of Information and Communications Technology, Keihanna*, 67–86. Stanford, CA: CSLI Publications. DOI: 10.21248/hpsg.2008.4.

Abstract

This paper describes a number of verbal argument marking patterns found in the world's languages and provides HPSG analyses for them. In addition to commonly-occurring variations of morphosyntactic alignment (e.g. nominative-accusative, ergative-absolutive), this paper also presents analyses of more complex phenomena, including ergativity splits, Austronesian-style focus-case systems, and direct-inverse systems and their interaction with case.

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 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, among others, mandatory sections on basic word order and basic lexical entries, and optional sections on sentential negation, coordination, and matrix yes/no questions. The lexicon section has recently been greatly enhanced, now allowing the description of complex inflectional morphology (O'Hara, 2008) and of an arbitary number of noun and verb classes.

This paper describes the implementation of a library that supports the marking of verbal arguments, principally via case. Development of such a library involves three steps. First, the typological range to be covered must be determined. Second, HPSG analyses must be developed for each of the possible marking systems. Finally, these analyses must be "factored" into a set of sub-analyses that the customization system can "snap together" in response to a linguist's answers to the

[†]The author would like to thank Emily Bender for her guidance, Dan Flickinger for help refining this paper's focus, Laurie Poulson for tense and aspect, Kelly O'Hara for morphology, Stefan Müller for calling my attention to TRALE, and Renée for proofreading. This work was supported by NSF grant BCS-0644097, a gift to the Turing Center by the Utilika Foundation, and the Max Planck Institute for Evolutionary Anthropology.

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

questionnaire and produce a consistent grammar. This paper will focus on the second step, the development of analyses, for several complex argument marking patterns, including split ergativity, focus-case marking, and direct-inverse languages, in which argument marking is sensitive to grammatical scales.

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 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.[2] 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.

Some languages have mandatory verbal arguments marked by additional cases beyond those marking intransitive subjects, agents, and patients. The Matrix customization questionnaire supports the description of an arbitrary number of additional case labels, which can then be used when describing the case of lexical items. In this paper, however, I will generally confine my attention only to cases marking the S, A, and O roles.

Nominative-accusative, ergative-absolutive, and tripartite NP case marking can be specified on verb lexical types using the ARG-ST feature (Manning and Sag, 1998) to constrain the argument structure, with the Argument Realization Principle providing the identities with the SUBJ and COMPS lists:[3]

(1) Nominative-Accusative

[2]There are nominative-accusative languages, including English and German, in which the nominative case only marks the S or A argument of finite verbs. Modeling the interaction of case-marking and verb form in the customization system is an area for future work.

[3]The current version of the system treats S and A as the subject and O as an object by placing them on the SUBJ and COMPS lists, respectively. In fact, this is not an adequate analysis cross-linguistically. Some languages show inter-clausal or syntactic ergativity, in which S and O pattern together in constructions including coordination and relative clauses (Dixon, 1979, 127). Manning (1996) describes an analysis of the variation between morphological and syntactic ergativity; however, the current version of the Matrix questionnaire includes almost no multi-clausal phenomena (the exception being coordination), so support for syntactic ergativity has been left for future work.

$$\begin{bmatrix} \textit{intransitive-verb-lex} \\ \textsc{synsem..head.val.subj}\left\langle \boxed{1} \right\rangle \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad \textit{nom} \right] \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{transitive-verb-lex} \\ \textsc{synsem..head.val} \begin{bmatrix} \textsc{subj} & \left\langle \boxed{1} \right\rangle \\ \textsc{comps} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad \textit{nom} \right], \boxed{2}\left[\textsc{..head.case} \quad \textit{acc} \right] \right\rangle \end{bmatrix}$$

(2) Ergative-Absolutive

$$\begin{bmatrix} \textit{intransitive-verb-lex} \\ \textsc{synsem..head.val.subj}\left\langle \boxed{1} \right\rangle \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad \textit{abs} \right] \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{transitive-verb-lex} \\ \textsc{synsem..head.val} \begin{bmatrix} \textsc{subj} & \left\langle \boxed{1} \right\rangle \\ \textsc{comps} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad \textit{erg} \right], \boxed{2}\left[\textsc{..head.case} \quad \textit{abs} \right] \right\rangle \end{bmatrix}$$

(3) Tripartite

$$\begin{bmatrix} \textit{intransitive-verb-lex} \\ \textsc{synsem..head.val.subj}\left\langle \boxed{1} \right\rangle \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad s \right] \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{transitive-verb-lex} \\ \textsc{synsem..head.val} \begin{bmatrix} \textsc{subj} & \left\langle \boxed{1} \right\rangle \\ \textsc{comps} & \left\langle \boxed{2} \right\rangle \end{bmatrix} \\ \textsc{arg-st}\left\langle \boxed{1}\left[\textsc{..head.case} \quad a \right], \boxed{2}\left[\textsc{..head.case} \quad o \right] \right\rangle \end{bmatrix}$$

The analysis of case in the Grammar Matrix case library also provides, in the

lexicon section of the questionnaire, two strategies for actually marking the case on the NP arguments: marking of whole NPs via case-marking adpositions, or morphological marking on nouns, determiners, or both.

2.1 Split Ergativity

Many languages are neither consistently ergative nor consistently accusative. Such languages are said to display SPLIT ERGATIVITY. In order to support this case pattern, the Matrix customization system must be able to create grammars in which more than one kind of marking, commonly the ergative and accusative patterns, coexist.

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

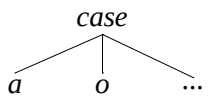
1. Semantic nature of the main verb
2. Semantic nature of the core NPs
3. Tense, aspect, or mood of the clause
4. Grammatical status of the clause

2.1.1 Semantic Nature of Main Verb

The first type of split occurs in two subtypes. In the first, 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 Split-S languages as having the following simple case hierarchy (the location of any additional cases in the hierarchy is represented by ...):

(4)



Based on this case type, Split-S grammars have a single transitive verb class with A- and O-marked arguments, but two intransitive verb classes:

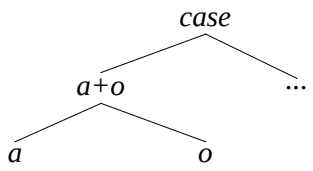
(5)
$$\begin{bmatrix} a\text{-}intrans\text{-}verb\text{-}lex \\ \text{ARG-ST} \left\langle \begin{bmatrix} \text{..HEAD.CASE} & a \end{bmatrix} \right\rangle \end{bmatrix}$$
$$\begin{bmatrix} o\text{-}intrans\text{-}verb\text{-}lex \\ \text{ARG-ST} \left\langle \begin{bmatrix} \text{..HEAD.CASE} & o \end{bmatrix} \right\rangle \end{bmatrix}$$

The questionniare allows the user/linguist to define verb lexical entries by defining any number of verb classes, each of which contains any number of stems. For each user-defined verb class, the user/linguist can choose which of the three lexical types above it derives from.

The other subtype is called FLUID-S. Fluid-S languages have, in addition to the two classes of verbs described above for Split-S languages, an additional intransitive verb class in which the single argument 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; however, it is possible to model the three intransitive verb classes.

I analyze Fluid-S languages with a slightly more articulated case hierarchy:

(6)



Fluid-S grammars include, in addition to the two lexical types above in 5, a lexical type for the fluid-marking verb class. This type simply specifies that the case of intransitive subjects is a supertype of both A and O:

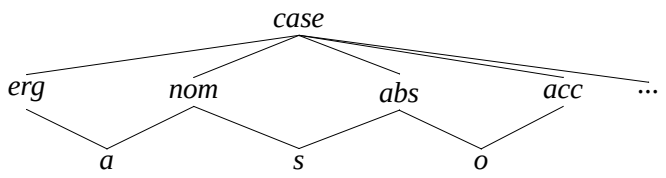
(7) $$\begin{bmatrix} a{+}o\text{-}intrans\text{-}verb\text{-}lex \\ \text{ARG-ST}\left\langle \left[\text{..HEAD.CASE} \quad a{+}o \right] \right\rangle \end{bmatrix}$$

2.1.2 Semantic Nature of NPs

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.

I analyze such a split with a rather more articulated case hierarchy:

(8)



For this type of language, the customization system will produce the same verb lexical types, shown in (3), that it would for a tripartite language. That is, an intransitive verb's sole argument is specified to take S case, while a transitive verb's agent and patient arguments take A and O, respectively. Then, when creating noun classes in the lexicon section of the questionnaire, the user/linguist will be prompted to specify for each class whether it is marked for *nom* (which unifies

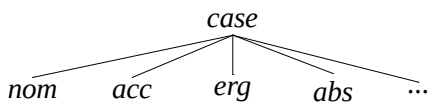
with *s* and *a*) and *acc* (which unifies just with *o*, or for *erg* (which unifies just with *a*) and *abs* (which unifies with *s* and *o*). This analysis puts the complexity in the right place in the lexicon for languages where the split is conditioned on the noun: verbs are not split, instead deriving from either the single intransitive or the single transitive type, while nouns are divided into classes based on whether they take the nominative-accusative or the ergative-absolutive pattern.

2.1.3 Clausal Splits

The third and fourth types of splits are both conditioned on clausal features. The third type is conditioned on the tense, aspect, or mood of the clause. 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.

The third and fourth types can be analyzed in the same way. The case hierarchy is flat, and has at least four values:

(9)



Verb lexical items have no case specified on their arguments; instead, a set of mandatory lexical rules is used to constrain the CASE values on their ARG-ST lists. For languages with the third type of split, the lexical rule that marks the conditioning feature (e.g. the past-tense morpheme) will constrain the CASE value of the arguments. For languages with the fourth type of split, two non-spelling-changing lexical rules can be used, along with the Matrix's MC (main clause) feature, to achieve the proper analysis: one rule marks the clause as [MC +] and constrains the cases on ARG-ST to one pattern, while the second rule marks the clause as [MC −] and constrains the cases on ARG-ST to the other pattern. However, at the time the case library was implemented, the Matrix customization system had no support for tense, aspect, or mood, nor for any phenomenon involving a subordinate clause, so there was no way to describe languages of the third or fourth type via the questionnaire.[4]

2.2 Focus-case Systems

Some Austronesian languages display an interesting variant of verbal argument marking (Comrie, 1989, 120). In Tagalog (Austronesian, Philippines), a language of this type, noun phrase arguments must be marked by one of several case-marking prepositions, one of which marks an NP as the FOCUS (Comrie, 1989, 121). The

[4]But see Poulson (forthcoming) for the details of a library for tense and aspect currently under development.

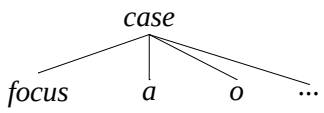
focus is marked by *ang*, while agent and patient are marked by *ng*. Every clause must have at least one argument marked as the focus. In intransitive clauses, this will be the sole argument. In transitive clauses, the verb is marked by one of a set of affixes that tell how the focus-marked NP should be interpreted, including among others agent-focus and patient-focus affixes. This pattern can be seen in the following examples (Comrie, 1989, 121):

(10) *Bumili* *ang* *babae* *ng* *baro*
bought-AGENT.FOCUS FOCUS woman PATIENT dress
'The woman bought a dress' [tgl]

(11) *Bimili* *ng* *babae* *ang* *baro*
bought-PATIENT.FOCUS AGENT woman FOCUS dress
'A/the woman bought the dress' [tgl][5]

This manner of argument marking is neither accusative nor ergative, instead constituting a distinct pattern. I analyze it as follows, using a slight modification of the analysis in §2. The case hierarchy is:

(12)



NPs are marked for agent, patient, or focus case, either directly in the lexicon or via case-marking adpositions. The sole argument on the ARG-ST of the intransitive verb lexical type is specified to have focus case. The lexical type of transitive verbs has an ARG-ST that is unspecified for case. In the lexicon section of the questionnaire, each type of focus-marking that can appear on a verb (including agent and patient focus) is implemented via a lexical rule that both applies the appropriate spelling change and constrains the cases of the arguments on ARG-ST. The rules for agent- and patient-focus marking are:

(13)
$$\left[\begin{array}{ll} \textit{agent-focus-verb-lex-rule} & \\ \text{INPUT} & \left\langle \boxed{1}, \textit{transitive-verb-lex} \right\rangle \\ & F_{af}(\boxed{1}), \\ \text{OUTPUT} & \left\langle \left[\text{ARG-ST} \left\langle \left[\ldots\text{CASE} \quad \textit{focus}\right], \left[\ldots\text{CASE} \quad \textit{patient}\right] \right\rangle \right] \right\rangle \end{array}\right]$$

[5]Comrie actually uses the terms *actor* and *undergoer*, but I use *agent* and *patient* here for consistency. Note that, although a single case-marker *ng* is used to mark both agents and patients in Tagalog, my analysis distinguishes between agent and patient, allowing it to model languages where they are marked differently.

$$\left[\begin{array}{ll} \textit{patient-focus-verb-lex-rule} & \\ \textsc{input} & \left\langle \boxed{1}, \textit{transitive-verb-lex} \right\rangle \\ \textsc{output} & \begin{array}{l} F_{pf}(\boxed{1}), \\ \left\langle \left[\textsc{arg-st} \left\langle \left[\ldots\textsc{case} \quad \textit{agent}\right], \left[\ldots\textsc{case} \quad \textit{focus}\right] \right\rangle \right] \right\rangle \end{array} \end{array}\right]$$

2.3 Direct-inverse Languages

In languages with DIRECT-INVERSE marking, the marking of verbal arguments is sensitive to a grammatical hierarchy. If the agent is ranked more highly on the hierarchy than the patient, then the clause is said to be DIRECT; if the patient is higher, the clause is said to be INVERSE. For a concrete example, let us consider the Algonquian languages, where the hierarchy is primarily sensitive to person:

(14) 2nd $>$ 1st $>$ 3rd proximate $>$ 3rd obviative

When a transitive clause contains two non-coreferential 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 usually "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.

(14) is often referred to in the literature as a *hierarchy*, but it differs markedly from the sort of multiply-inheriting type hierarchies used in HPSG. The hierarchy in (14) only implies one-dimensional precedence relationships among the positions on the hierarchy; in contrast, HPSG-style type hierarchies involve arbitrary pairwise inheritance relationships among the items they contain. To avoid confusion, I will hereafter refer to grammatical hierarchies like (14) as SCALES.[6]

The following examples from Fox (Algonquian, North America) illustrate how argument marking works in a direct-inverse language (Comrie, 1989, 129):

(15) *ne -waapam-aa -wa*
1SG see-DIRECT 3
'I see him.' [sac]

(16) *ne -waapam-ek -wa*
1SG see-INVERSE 3
'He sees me.' [sac]

[6]The usage of *hierarchy* to refer to such scales, 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).

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 constrain the verb's arguments differently for direct and inverse clauses. It would be convenient when modeling this aspect of direct-inverse languages (via lexical rules, say) if there were a formal mechanism for stating scale constraints compactly, perhaps something like:

(17)
$$\left[\begin{array}{ll} \textit{direct-verb-lex-rule} & \\ \text{INPUT} & \left\langle \boxed{1}, \ldots \right\rangle \\ \text{OUTPUT} & \left\langle F_{dv}(\boxed{1}, \left[\text{ARG-ST} \left\langle \boxed{2}, \boxed{3} \right\rangle\right]) \right\rangle \end{array}\right] \& \boxed{2} >> \boxed{3}$$

$$\left[\begin{array}{ll} \textit{inverse-verb-lex-rule} & \\ \text{INPUT} & \left\langle \boxed{1}, \ldots \right\rangle \\ \text{OUTPUT} & \left\langle F_{iv}(\boxed{1}, \left[\text{ARG-ST} \left\langle \boxed{2}, \boxed{3} \right\rangle\right]) \right\rangle \end{array}\right] \& \boxed{2} << \boxed{3}$$

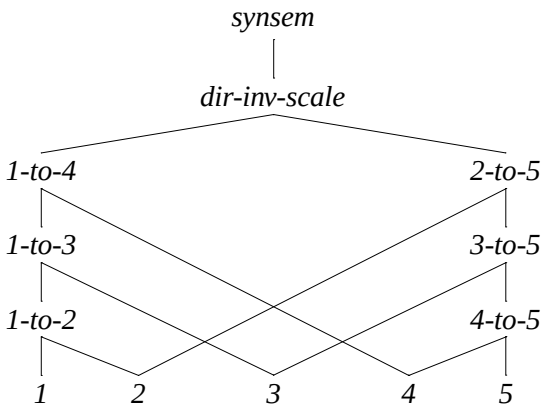
However, no such mechanism is available to us, so another method of analyzing scales is required.[7] It would be possible, of course, to simply create a lexical rule for each possible pair of positions on the scale, but this would mean having on the order of n^2 lexical rules for an n-position scale. It would be better to somehow model the scale with a type hierarchy.

Perhaps, noticing that it is necessary to address ranges of the scale that start at the left or the right end, we might try to model the scale using a type hierarchy like (18) (labeling the positions on the scale from 1 through 5), which is then used to constrain the series of lexical rules in (19) (which all derive from a single rule that applies the direct morphology to the verb):[8]

[7]Note, however, that other systems for implementing HPSG grammars are more powerful. In particular, the TRALE system (Meurers et al., 2002) can state constraints like those in (17) using its complex antecedent feature (Stefan Müller, personal communication, October 2008).

[8]This analysis models scales using subtypes of *synsem*, anticipating that the features involved may be syntactic or semantic. It is possible that a more specific feature structure would do (e.g. *local* or something within *cat* or *cont*), in some or all languages. This is left for future work.

(18)



(19)

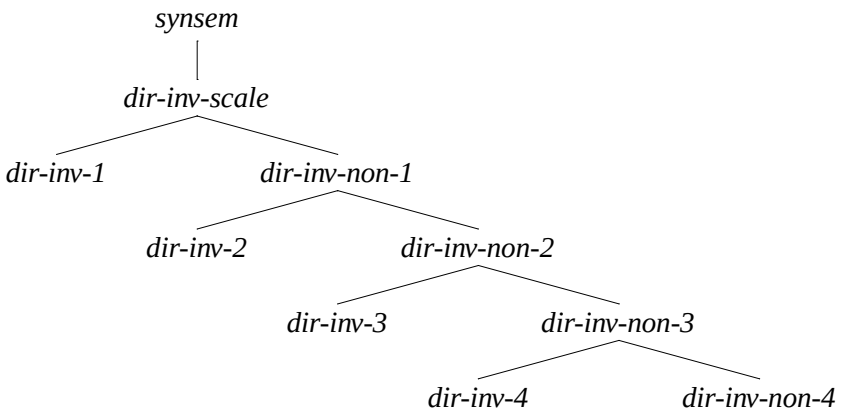
$$\begin{bmatrix} \textit{direct-verb-lex-rule-1} \\ \text{ARG-ST}\left\langle \textit{1}, \textit{2-to-5} \right\rangle \end{bmatrix} \qquad \begin{bmatrix} \textit{direct-verb-lex-rule-2} \\ \text{ARG-ST}\left\langle \textit{1-to-2}, \textit{3-to-5} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{direct-verb-lex-rule-3} \\ \text{ARG-ST}\left\langle \textit{1-to-3}, \textit{4-to-5} \right\rangle \end{bmatrix} \qquad \begin{bmatrix} \textit{direct-verb-lex-rule-4} \\ \text{ARG-ST}\left\langle \textit{1-to-4}, \textit{5} \right\rangle \end{bmatrix}$$

Unfortunately, this set of rules would produce spurious ambiguity when applied to some sentences. While a sentence with, say, a subject from class 1 and an object from class 2 would parse just 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 parse four times, once for each of the above rules.

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

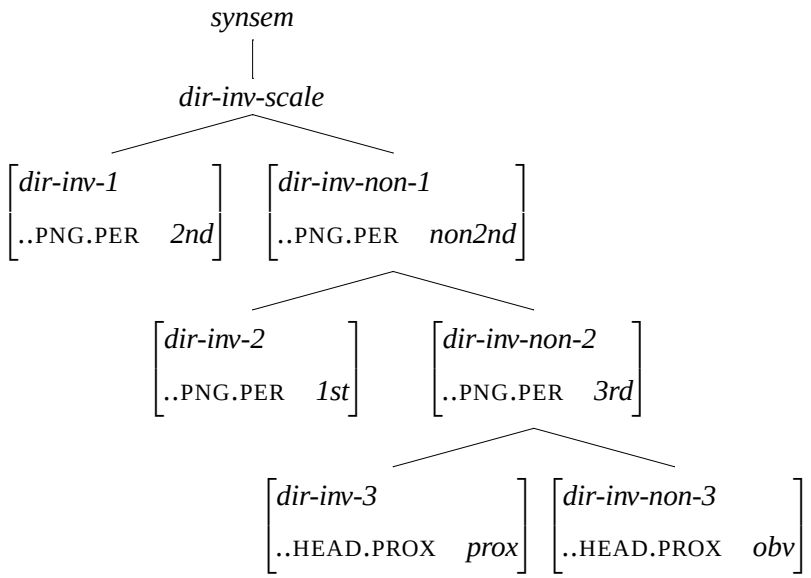
(20)



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 a

concrete example, below are the type hierarchy (21) and lexical rules (22) for an Algonquian language with the scale in (14):

(21)



(22)

$$\begin{bmatrix} \textit{direct-verb-lex-rule-1} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-1}, \textit{dir-inv-non-1} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-1} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-1}, \textit{dir-inv} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{direct-verb-lex-rule-2} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-2}, \textit{dir-inv-non-2} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-2} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-2}, \textit{dir-inv-2} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{direct-verb-lex-rule-3} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-3}, \textit{dir-inv-non-3} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-3} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-3}, \textit{dir-inv-3} \right\rangle \end{bmatrix}$$

A further set of lexical rules that are sensitive to the value of the DIRECTION feature are defined by the user/linguist in the lexicon section of the questionnaire. These rules actually apply whatever spelling changes are associated with the direct and inverse forms of the verb; for example, handling the Fox examples in (15) and (16) would require a direct-marking rule for the suffix *-aa* and an inverse-marking rule for the suffix *-ek*. It would be possible in principle to merge the scale-constraining rules like those in (22) and the rules marking direct or inverse on the verb into a single paradigm of lexical rules; however, the questionnaire allows any number of morphological "slots" to be created that are sensitive to the

DIRECTION feature, raising the question of which slot's rules should also specify the constraints in (22). To avoid this issue, the customization system always separates the scale-constraining rules from any lexical rules that implement user-defined verb morphology.

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.

Note that this analysis does not allow the parsing of transitive sentences where both NP arguments occupy the same position on the scale. This is correct for at least some Algonquian languages including Nishnaabemwin, where coreferential NP arguments require a reflexive form and two third person arguments can be distinguished using the obviative (Valentine, 2001, 273). Another possibility, languages where both NP arguments may occupy the same position on the scale, is analyzed below in §2.4.

It is worth noting some drawbacks to this analysis. First, it requires, for a scale with n positions, $2(n-1)$ lexical rules. Furthermore, the type hierarchy in (21) is only arbitrarily right-branching. An 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.

Finally, it should be noted that the leaf types in the *dir-inv-scale* hierarchy, which are certainly necessary because they encode the positions on the grammatical scale, need not be arranged in a single hierarchy in order to model the language. The leaves could all be independent subtypes of *synsem*, and the verb lexical rules could be stated in exactly the same way without a *dir-inv-scale* supertype. However, there is a good reason to prefer a hierarchy to independent types. In (21), the features of the types *dir-inv-2* and *dir-inv-non-2* had better be compatible with those of *dir-inv-non-1*—otherwise, the latter type cannot be opposed with *dir-inv-1* in verb argument structures to distinguish NPs at the left of the scale from NPs at any position further down the scale. Since software systems can contain bugs, it is therefore valuable, as a "sanity check" on grammars produced by the customization system, to arrange the leaf types into a hierarchy. If the types are not compatible, loading the grammar with the LKB will produce an error rather than apparently succeeding but parsing and generating incorrectly. In other words, it ought to be possible to arrange the types encoding the grammatical scale into a hierarchy, and in fact, the grammar is seriously inconsistent if they cannot be so arranged, so to be safe, the customization system does so.

2.4 Fore

Scales can also control the verbal argument marking patterns in languages that lack direct or inverse marking on the verb. One such language is Fore (Trans-New

Guinea, Papua New Guinea), where the relative position of agent and patient on a scale correlates with the presence or absence of a marker on the agent NP. The scale governing argument marking in Fore is:

(23) pronoun, name, kin term > human > animate > inanimate

The operation of this hierarchy can be seen in the following examples (Scott 1978, 116, Blake 2001, 122):

(24) *yaga: wá aegúye*
pig man 3SG.hit.3SG
'The man kills the pig' [for]

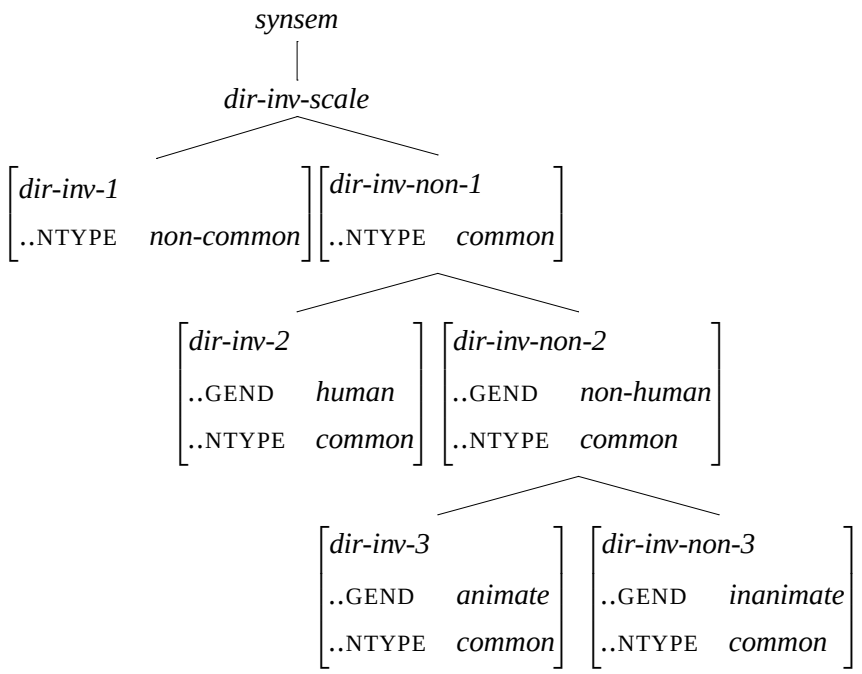
(25) *yaga:-wama wá aegúye*
pig-DLN man 3SG.hit.3SG
'The pig kills the man' [for]

(26) *wa yága:-wama aegúye*
man pig-DLN 3SG.hit.3SG
'The pig kills the man' [for]

An extra suffix *-wama* (which Scott (1978) describes as a "delineator") appears on the agent when it is lower on the hierarchy than the patient. Scott describes these facts of Fore without referring to it as a direct-inverse language; however, I will show that this marking pattern can be analyzed by treating Fore as direct-inverse language where, instead of marking on the verb, it is the marking of case on NPs that is sensitive to direct or inverse clauses.

I analyze Fore as an ergative-absolutive language, where ergative is marked by the delineator suffix *-wama*. To capture the distinction between types and genders of nouns, nominal heads have an additional NTYPE feature with the values *common* and *non-common*, and the GEND feature on PNG under INDEX has the values *human*, *non-human*, *animate*, and *inanimate* (where the latter two are subtypes of *non-human*). The *dir-inv-scale* hierarchy in the grammar is:

(27)



The grammar also contains a set of constant verb lexical rules, one of which will apply to the verb in each transitive clause, constraining the items on its ARG-ST list:

(28)

$$\begin{bmatrix} \textit{direct-verb-lex-rule-1} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-1}, \textit{dir-inv-scale} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-1} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-1}, \textit{dir-inv} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{direct-verb-lex-rule-2} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-2}, \textit{dir-inv-non-1} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-2} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-2}, \textit{dir-inv-2} \right\rangle \end{bmatrix}$$

$$\begin{bmatrix} \textit{direct-verb-lex-rule-3} \\ \text{..HEAD.DIRECTION } \textit{dir} \\ \text{ARG-ST} \left\langle \textit{dir-inv-3}, \textit{dir-inv-non-2} \right\rangle \end{bmatrix} \begin{bmatrix} \textit{inverse-verb-lex-rule-3} \\ \text{..HEAD.DIRECTION } \textit{inv} \\ \text{ARG-ST} \left\langle \textit{dir-inv-non-3}, \textit{dir-inv-3} \right\rangle \end{bmatrix}$$

Compare the ARG-ST constraints in the rules in (28) with those in (22). The inverse rules are similar, but notice that the direct rules for Fore, rather than constraining agents and patients using types from the same level in the hierarchy, instead constrain patients to types that are the supertypes of their corresponding agents. For example, in *direct-verb-lex-rule-1*, *dir-inv-1* is opposed with *dir-inv-scale* rather than with *dir-inv-non-1*. This is necessary because Fore, unlike the

Algonquian languages described in §2.3, allows clauses where both arguments occupy the same position on the scale (Scott, 1978, 115).[9] The customization system allows the description of both types of languages in its questionnaire.

After one of the above rules has applied to a verb stem, another constant verb lexical rule from the set below applies. These rules are sensitive to the value of the DIRECTION feature and constrain the case of the verb's arguments appropriately.

(29)
$$\left[\begin{array}{ll} \textit{direct-lex-rule} & \\ \text{..HEAD.DIRECTION} & \textit{dir} \\ \text{..VAL.SUBJ} & \left\langle \left[\text{..HEAD.CASE } \textit{abs} \right] \right\rangle \\ \text{..VAL.COMPS} & \left\langle \left[\text{..HEAD.CASE } \textit{abs} \right] \right\rangle \end{array}\right]$$

$$\left[\begin{array}{ll} \textit{inverse-lex-rule} & \\ \text{..HEAD.DIRECTION} & \textit{inv} \\ \text{..VAL.SUBJ} & \left\langle \left[\text{..HEAD.CASE } \textit{erg} \right] \right\rangle \\ \text{..VAL.COMPS} & \left\langle \left[\text{..HEAD.CASE } \textit{abs} \right] \right\rangle \end{array}\right]$$

Note that constraints on the rules in (28) and (29) could have been folded into a single paradigm of rules by having the direct rules derive from *direct-lex-rule* and the inverse rules from *inverse-lex-rule*. However, because this analysis of Fore treats it as a direct-inverse language, the structure of the lexical rule system produced by the customization system parallels that in §2.3 above, with separate two sets of rules, one implementing scale constraints and the other marking clauses as direct or inverse (via verb morphology in Algonquian and via case-marking in Fore).

3 Results

In order to test the direct-inverse section of the customization system, I have filled out the questionnaire and created two small grammars, one for a language fragment resembling an Algonquian language and the other for a fragment resembling Fore. Below, I show the coverage of each grammar on a suite of sentences designed to test correct parsing.

[9] The delineator in Fore can also be used to make available dispreferred word orders with scale-equivalent arguments, but the current version of the customization system is not powerful enough to capture such an interaction between word order and argument marking. This grammatical fact must therefore be left for future work.

3.1 Pseudo-Algonquian

The Algonquian languages have direct and inverse marking of the verb, controlled by the scale in (14), repeated here for convenience:

(30) 2nd > 1st > 3rd proximate > 3rd obviative

To demonstrate the Matrix customization system's ability to handle such languages, a simple pseudo-Algonquian grammar was created via the questionnaire. It has no case marking; an additional head feature called PROXIMITY, used to mark proximate and obviative forms of third-person nouns; SVO word order[10]; and the scale in (30).

The pseudo-Algonquian lexicon contains a transitive verb *tv* and the nominal forms *1P*, *2P*, and *3P*, which have lexically-specified values of PERSON. Verbs take one of two suffixes: *-DIR*, which marks direct form, and *-INV*, which marks the inverse. Third person nouns take one of two suffixes: *-PROX* for the proximate or *-OBV* for obviative.

The grammar produces the judgments marked on the sentences below:

(31)	2P tv-DIR 1P	*2P tv-INV 1P
	2P tv-DIR 3P-PROX	*2P tv-INV 3P-PROX
	2P tv-DIR 3P-OBV	*2P tv-INV 3P-OBV
	1P tv-DIR 3P-PROX	*1P tv-INV 3P-PROX
	1P tv-DIR 3P-OBV	*1P tv-INV 3P-OBV
	3P-PROX tv-DIR 3P-OBV	*3P-PROX tv-INV 3P-OBV
	3P-OBV tv-INV 3P-PROX	*3P-OBV tv-DIR 3P-PROX
	3P-OBV tv-INV 1P	*3P-OBV tv-DIR 1P
	3P-OBV tv-INV 2P	*3P-OBV tv-DIR 2P
	3P-PROX tv-INV 1P	*3P-PROX tv-DIR 1P
	3P-PROX tv-INV 2P	*3P-PROX tv-DIR 2P
	1P tv-INV 2P	*1P tv-DIR 2P

These sentences are divided into four groups. Those in the upper-left quadrant are grammatical because the agent (first argument) outranks the patient and the verb is in direct form. Those in the lower-left quadrant are grammatical because the patient outranks the agent and the verb is in the inverse form. The sentences in the right column have the same arguments as those on the left, but *-DIR* and *-INV* have been reversed, so they are all ungrammatical.

3.2 Pseudo-Fore

The pseudo-Fore grammar has ergative-absolutive case marking; human, animate, and inanimate genders; an additional head feature called NTYPE that distinguishes

[10] Algonquian languages typically have free word order, but to make it easier to create both grammatical and ungrammatical test sentences, this pseudo-Algonquian is constrained to be SVO.

pronouns, names, kin terms, and common nouns; verb-final word order; and the scale in (23), repeated here for convenience:

(32) pronoun, name, kin term > human > animate > inanimate

The pseudo-Fore lexicon contains a transitive verb *tv* and the nouns *pro* (a pronoun), *human*, *anim*, and *inanim*, the latter three being common nouns of the obvious gender. The only inflection is the *-ERG* suffix on nouns.

The grammar produces the judgments marked on the sentences below:

(33)			
	pro pro tv	*pro pro-ERG tv	*pro-ERG pro tv
	pro human tv	pro human-ERG tv	*pro-ERG human tv
	pro anim tv	pro anim-ERG tv	*pro-ERG anim tv
	pro inanim tv	pro inanim-ERG tv	*pro-ERG inanim tv
	human human tv	*human human-ERG tv	*human-ERG human tv
	human anim tv	human anim-ERG tv	*human-ERG anim tv
	human inanim tv	human inanim-ERG tv	*human-ERG inanim tv
	anim anim tv	*anim anim-ERG tv	*anim-ERG anim tv
	anim inanim tv	anim inanim-ERG tv	*anim-ERG inanim tv
	inanim inanim tv	*inanim inanim-ERG tv	*inanim-ERG inanim tv

Sentences in the left column are all grammatical because no case is marked—in fact, the sentences with both arguments from the same scale position (e.g. *pro pro tv*, *human human tv*) are ambiguous and parse twice due to Fore's verb-final word order. The sentences in the center column have the second argument, which is always of lower or equal scale rank, marked with the ergative suffix. They are grammatical except where the two arguments are of equal rank, in which case Fore does not allow the ergative. The sentences in the third column have the first argument, which is always of higher or equal scale rank, marked with the ergative suffix. They are all ungrammatical because ergative may only be marked on the lower-ranked argument.

4 Conclusion

In this paper I have described analyses of a number of verbal argument marking patterns. These included several case patterns: nominative-accusative, ergative-absolutive, tripartite, split ergative, and focus-case. I also described an analysis of direct-inverse languages, whose marking pattern was challenging to describe compactly in HPSG.

The development and implementation of such sets of analyses, where each analysis must be designed so that it can be plugged into an automatically-created Matrix-based grammar, represents an instance of what could be called computational linguistic typology. Rather than analyzing linguistic phenomena deeply but separately, as syntacticians often do, or collecting shallow descriptions of the range

a phenomenon in the world's languages, as typologists do, I instead analyze in detail the whole typological range of a phenomenon (here, verbal argument marking) within a single consistent framework. The resulting analyses are made available via the Matrix customization system, which emits grammars whose correctness can be verified against suites of test sentences.

The aim of this style of analysis is to bring to light unrecognized commonalities among human languages. This effort has already born some fruit. I have shown here that an analysis of direct-inverse languages based on a complex of lexical rules can be extended to other languages whose argument marking is conditioned on grammatical scales. I expect that the implementation of libraries for other linguistic phenomena for the Grammar Matrix will reveal further generalizations.

A more detailed presentation of the work described here, along with additional Matrix libraries for person, number, gender, and agreement, will form the core of my dissertation (Drellishak, forthcoming).

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 & Linguistic Typology, Second Edition*. Chicago: University of Chicago.

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.

Dixon, R. M. W. 1979. Ergativity. *Language* 55, 59–138.

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

Drellishak, Scott. forthcoming. *Widespread, but Not Universal: Improving the Typological Coverage of the Grammar Matrix*. Ph. D.thesis, University of Washington.

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, Edward L. and Comrie, Bernard. 1977. Noun Phrase Accessibility and Universal Grammar. *Linguistic Inquiry* 8, 63–99.

Manning, Christopher D. 1996. *Ergativity*. Stanford: CSLI.

Manning, Christopher D. and Sag, Ivan A. 1998. Argument Structure Valence and Binding. *Nordic Journal of Linguistics* 21, 107–144.

Meurers, W. Detmar, Penn, Gerald and Richter, Frank. 2002. A Web-based Instructional Platform for Constraint-Based Grammar Formalisms and Parsing. In Dragomir Radev and Chris Brew (eds.), *Effective Tools and Methodologies for Teaching NLP and CL*, pages 18–25, New Brunswick, NJ: The Association for Computational Linguistics.

O'Hara, Kelly. 2008. *A Morphotactic Infrastructure for a Grammar Customization System*. Masters Thesis, University of Washington.

Poulson, Laurie. forthcoming. *Grammatical Tense and Aspect: A Cross-Linguistic Implementation*. Ph. D.thesis, University of Washington.

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

Valentine, J. Randolph. 2001. *Nishnaabemwin Reference Grammar*. Toronto: University of Toronto Press.