Complex Case Phenomena in the Grammar Matrix

Scott Drellishak
University of Washington

July 28, 2008

[1]This work is supported by NSF grant BCS-0644097, a gift to the Turing Center by the Utilika Foundation, and the Max Planck Institute for Evolutionary Anthropology

The LinGO Grammar Matrix (Bender et al., 2002)

- Distill the wisdom of existing broad-coverage grammars
- Provide a typologically-informed foundation for building grammars of natural languages in software
- Syntax-semantics interface consistent with HPSG and Minimal Recursion Semantics (Copestake et al., 2005)
- Today I'll describe the implementation of case and verbal argument marking in the Matrix
- I'll discuss in detail direct-inverse languages, which have an argument-marking strategy particularly challenging to analyze with the tools we use

Matrix Libraries

- Matrix intended to cover all languages, but there exist phenomena that are widespread but not universal
- If not universal, do they belong in the Matrix?
- Solution: divide the Matrix into:
  - The universal or "core" Matrix
  - Matrix "libraries" covering non-universal phenomena
- Libraries are exposed to the user-linguist via a typological questionnaire: http://www.delph-in.net/matrix/customize/matrix.cgi
- Based on answers, we customize an HPSG grammar expressed in TDL (type description language) and compatible with the LKB (Copestake, 2002)

Case

- Part of my recent work has been a library for case
- CASE is "a system of marking dependent nouns for the type of relationship they bear to their heads." (Blake, 2001)
- A little broader: marked on noun phrases, on one or more of: nouns, pronouns, determiners, adjectives
- I also take it to include NPs marked by adpositions, though not everyone does
- Extremely complex phenomenon; this library only covers case-marking on the selected arguments of verbs, up to two per verb
  - Narrowing the range of phenomena simplifies the implementation
  - Excludes, e.g., noun-modifier case concord, oblique case, and possessive cases (unless used to mark a verbal argument)

Morphosyntactic Alignment

- Covering verbal arguments means dealing with morphosyntactic alignment
- That is, how a language marks three key roles:
  - **Subject**, the sole argument of intransitives (S)
  - **Agent**, the actor argument of transitives (A)
  - **Patient**, the acted upon argument of transitives (O or P)
- There are many different patterns cross-lingusitically

Cross-linguistic Variation

- (Using the terminology of Dixon (1994))
- **Nominative-accusative**: S marked like A, O different
  ex: English, Japanese
- **Ergative-absolutive**: S marked like O, A different
  ex: (almost) Dyirbal (Australian, Pama-Nyungan)
- **Tripartite**: S, A, and O all marked differently
  ex: Wangkumara (Australian, Pama-Nyungan)
- **Split-S**: S is marked like A for some intransitives, like O for others
  ex: Mandan (Siouan)
- **Fluid-S**: S is marked like A for some intransitives, like O for others, and like either A or O for still others
  ex: Tsova-Tush (N.E. Caucasian)

Cross-linguistic Variation

- Pure ergative-absolutive languages are very rare; most show **split ergativity**
- One part of the grammar shows a nominative-accusative pattern, while another shows an ergative-absolutive pattern
- Two major kinds: (my abbreviations)
  - **Split-N**: Some nominals have nominative-accusative marking, while others have ergative-absolutive
    ex: Dyirbal (nom-acc in 1st and 2nd persons)
  - **Split-V**: Based on some feature (e.g. tense, aspect, mood, main/subordinate status) of the verb, arguments take either ergative-absolutive or nominative-accusative arguments
    ex: many Indo-Iranian languages, inc. Hindi/Urdu

Cross-linguistic Variation

- The **Focus case** marking pattern appears in some Austronesian languages
- One argument is marked by a case (sometimes called focus) whose role is assigned by a morpheme on the verb:
- Tagalog:

(1)

Bumili	*ang*	*babae*	*ng*	*baro*
bought-AGENT-FOC	FOC	woman	PATIENT	dress

'The woman bought a dress'

(2)

Bimili	*ng*	*babae*	*ang*	*baro*
bought-PATIENT-FOC	AGENT	woman	FOC	dress

'A/the woman bought the dress' (Comrie, 1989, 121)

Scale-sensitive Marking

- Other kinds of argument marking exist
- In **direct-inverse languages** (e.g. Algonquian languages), marking is sensitive to a scale
- NPs are ranked according to their naturalness as an agent
  - If A outranks O, verb in **direct** form
  - If O outranks A, verb in **inverse** form
- Other languages (e.g. Fore, Papuan) have case marking that's scale-sensitive
- Not generally analyzed as direct-inverse, but I show here that the same analysis can be used for both

Case Questionnaire

- (Demo)

Direct-inverse Languages

- Marks arguments, but is it case?
- When I started, I intended to treat it as case, but that's wrong
- Languages can have case and inverse marking in various combinations:
  - Direct-inverse verbs and case (Sahaptin, Penutian)
  - Direct-inverse verbs but no case (Algonquian)
  - Case-marking that's sensitive to a scale, but no overt marking of direct or inverse (Fore)
- So direct-inverse isn't case, but still falls under the slightly broader umbrella of verbal argument marking

A Challenge

- The LKB implements HPSG with multiple inheritance and unification
- Direct-inverse languages present a challenge
- It would be nice to do something like:

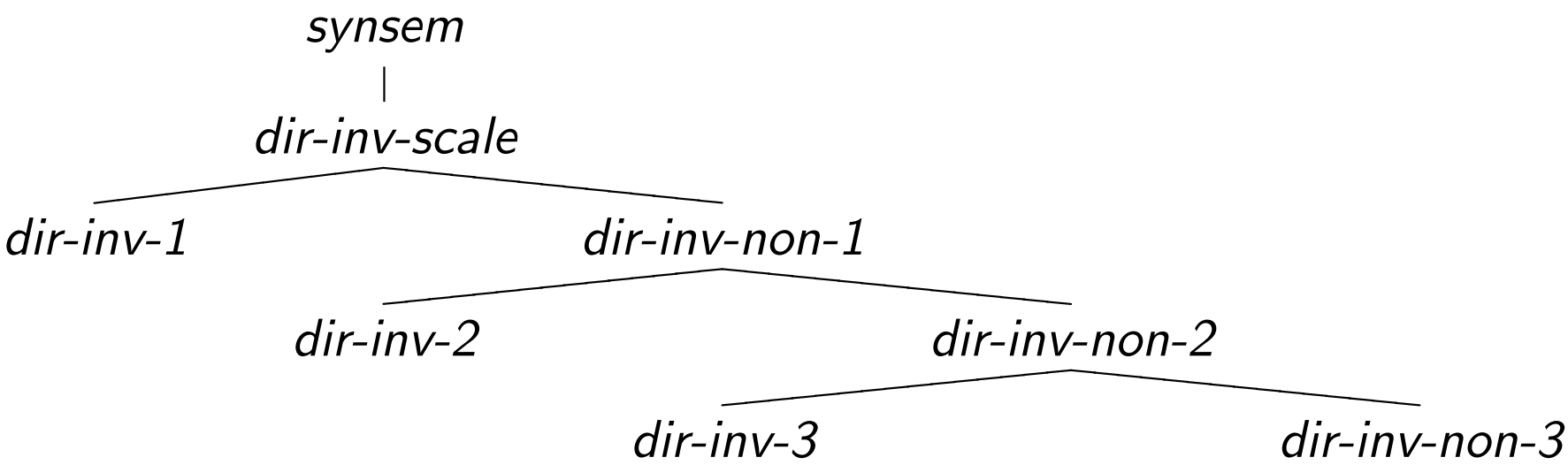
$$\left[\begin{array}{l} \textit{direct-verb-lex-rule} \\ \textsc{input} \left\langle \boxed{1}, \ \ldots \right\rangle \\ \textsc{output} \left\langle F_{dv}(\boxed{1}), \left[\textsc{arg-st} \left\langle \boxed{2}, \ \boxed{3} \right\rangle\right] \right\rangle \end{array}\right] \ \& \ \boxed{2} > \boxed{3}$$

- However, no such mechanism is available
- How can we analyze the direct-inverse pattern without specifying n^2 rules for n scale entries?

My Analysis

- Consider the scale for the Algonquian languages:
  2nd pers. $>$ 1st pers. $>$ 3rd pers. prox. $>$ 3rd pers. obv.
- To implement such a scale, I use two mechanisms:
  - A binary-branching hierarchy encoding the scale
  - A set of mandatory lexical rules

Scale Hierarchy



- Types on the left specify the features for a scale entry
- Types on the right specify features covering the rest of the scale

Algonquian Scale

- So, for Algonquian:

$$\begin{bmatrix} \textit{dir-inv-1} \\ \ldots\textsc{per} \quad \textit{2nd} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-1} \\ \ldots\textsc{per} \quad \textit{non-2nd} \end{bmatrix}$$

$$\begin{bmatrix} \textit{dir-inv-2} \\ \ldots\textsc{per} \quad \textit{1st} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-2} \\ \ldots\textsc{per} \quad \textit{3rd} \end{bmatrix}$$

$$\begin{bmatrix} \textit{dir-inv-3} \\ \ldots\textsc{per} \quad \textit{3rd} \\ \ldots\textsc{proximity} \quad \textit{proximate} \end{bmatrix} \qquad \begin{bmatrix} \textit{dir-inv-non-3} \\ \ldots\textsc{per} \quad \textit{3rd} \\ \ldots\textsc{proximity} \quad \textit{obviative} \end{bmatrix}$$

Lexical Rules

- For Algonquian, lexical rules for direct specifying:

  ARG-ST $\left\langle \textit{dir-inv-1}, \textit{dir-inv-non-1} \right\rangle$

  ARG-ST $\left\langle \textit{dir-inv-2}, \textit{dir-inv-non-2} \right\rangle$

  ARG-ST $\left\langle \textit{dir-inv-3}, \textit{dir-inv-non-3} \right\rangle$

- ...and three rules for inverse:

  ARG-ST $\left\langle \textit{dir-inv-non-1}, \textit{dir-inv-1} \right\rangle$

  ARG-ST $\left\langle \textit{dir-inv-non-2}, \textit{dir-inv-2} \right\rangle$

  ARG-ST $\left\langle \textit{dir-inv-non-3}, \textit{dir-inv-3} \right\rangle$

- (Hierarchy not strictly necessary–it's a sanity check on grammars)

Equal Agent and Patient

- What if agent and patient are ranked the same? Two possibilites:
  1. Impossible. Algonquian languages work this way. They have a reflexive form and proximate/obviative marking for 3rd.
  2. Possible. Fore works this way. Direct form is used, and then word order determines A and O.
- Analyzing languages of the second type requires only adjusting which *dir-inv-scale* subtypes appear in which rules:

ARG-ST $\left\langle \textit{dir-inv-1}, \textit{dir-inv-scale} \right\rangle$

ARG-ST $\left\langle \textit{dir-inv-2}, \textit{dir-inv-non-1} \right\rangle$

ARG-ST $\left\langle \textit{dir-inv-3}, \textit{dir-inv-non-2} \right\rangle$

ARG-ST $\left\langle \textit{dir-inv-non-3}, \textit{dir-inv-non-3} \right\rangle$

Example: Algonquian

- Group of North American languages with similar marking pattern
- Sensitive to the following scale (repeated):
  2nd pers. $>$ 1st pers. $>$ 3rd pers. prox. $>$ 3rd pers. obv.
- Obviation: one third person NP is proximate, the rest obviative
- The proximate NP is the topic of the discourse, usually the focus of the speaker's empathy, often the viewpoint character in a narrative (Dahlstrom, 1991, 91)
- The grammar shown here is simplified to focus on direct-inverse, and to have some ungrammatical examples:
  - Agent and patient agreement on the verb omitted
  - Word order fixed to SVO

Algonquian Results

- Results from a customized Algonquian grammar:

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

Algonquian Semantics (for "2P tv-DIR 1P")

```
[ LTOP: h1
  INDEX: e2 [ SF: PROP-OR-QUES ...  ]
  RELS <
    [ "_pronoun_n_rel"
      LBL: h3
      ARG0:  x4 [ x SORT: SEMSORT PNG.PER: 2ND ...  ]]
    [ "unspec_q_rel"
      ...  ]
    [ "_tv_v_rel"
      LBL: h1
      ARG0:  e2
      ARG1:  x4
      ARG2:  x8 [ x SORT: SEMSORT PNG.PER 1ST ...  ]]
    [ "_pronoun_n_rel"
      LBL: h9
      ARG0:  x8 ]
    [ "unspec_q_rel"
      ...  ]  >

  HCONS < h6 qeq h3 h11 qeq h9 > ]
```

Fore

- Fore (Papuan) is an interesting case: no direct or inverse form of the verb
- But has case marking sensitive to a hierarchy:
  pronoun, personal name, kin term $>$
  human $>$ animate $>$ inanimate
- When agent outranks patient, NPs are unmarked
- When patient outranks agent, agent is marked
- Case called "ergative" by (Blake, 2001, 122), but simply a "delineator" by Scott (1978)

Fore Results

- Results from a customized Fore grammar:

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

Fore Semantics (for "inanim-ERG pro tv")

```
[ LTOP: h1
  INDEX: e2 [ SF: PROP-OR-QUES ...  ]
  RELS <
    [ "_inanim_n_rel"
      LBL: h3
      ARG0:  x4 [ x SORT: SEMSORT ...  ]]
    [ "unspec_q_rel"
      ...  ]
    [ "_pronoun_n_rel"
      LBL: h8
      ARG0:  x9 [ x SORT: ...  ]]
    [ "unspec_q_rel"
      ...  ]
    [ "_tv_v_rel"
      LBL: h1
      ARG0:  e2
      ARG1:  x4
      ARG2:  x9 ]

  HCONS < h6 qeq h3 h11 qeq h8 > ]
```

Future Work

- Further argument-marking patterns, including word-order shifts
- Ditransitives
- Syntactic ergativity
- Huge variety of other non-argument-marking case phenomena
- Other features:
  person and number are in, but need more elaborate treatment
- For all phenomena, factorable analyses

Summary

- Wide variety of argument-marking case patterns as a Matrix library
- Direct-inverse argument marking is a challenge to implement with only multiple inheritance and unification
- I've shown an analysis that works, but requires multiple rules rather than a simple rule or principle that can be stated once
- Just scratched the surface of the customization system. Either try it out online or ask me for a demo.
  http://www.delph-in.net/matrix/customize/matrix.cgi
  http://students.washington.edu/sfd/

References

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. 1994. *Ergativity*. Cambridge: Cambridge University Press.

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