What's in a Word?

Refining the Morphotactic Infrastructure in the LinGO Grammar Matrix Customization System

Michael Wayne Goodman, Emily M. Bender

{goodmami,ebender}@uw.edu

University of Washington, Seattle, WA, USA

1. Background

Project Goals: Quick grammar customization through a typologically informed questionnaire

- The Grammar Matrix and its customization system are a repository of linguistic analyses for typologically common phenomena (Bender et al., 2002, 2010).
  - Eases burden on grammar developer.
  - Promotes comparability among grammars.
- Core grammar contains types common to all languages.
- Language-specific types customized in questionnaire.

Morphological Paradigm: Inferential and incremental.

- Using Stump's (2001) terms, our system is **inferential** and **incremental**.
- Stump argues for a realizational model because of multiple exponence and zero realization, but they are not problematic for us.
  - Syntactic and semantic contributions of morphemes are modeled by unification.
  - All constraints must be modeled—if some do not have an overt morpheme, a zero-marked rule is used.

Morphotactics: Lexical rule interaction.

- For this work, we are concerned with **morphotactics**, i.e., the co-occurrence restrictions of morphemes.
- Separate from both syntactic and phonological (orthographic) consequences of morphemes.
  - We delegate non-concatenative phonological effects to an external morphonological processor.
  - Syntactic/semantic constraints applied by lexical rules.
- Lexical rules can **require** or **forbid** other lexical rules. Otherwise they are optional.

2. Implementation

Slots and Morphemes: Defining lexical rules.

- Slots are akin to morphological paradigms.
  - Define where morphemes occur.
  - Place constraints on other morphemes.
- With a slot, specific morphemes are implemented as lexical rules, constrained to provide:
  - Syntactic contribution (if any).
  - Orthographic contribution (if any).

INFLECTED: Bool and Beyond.

- The feature INFLECTED defines whether a lexeme can be used in a phrase.
- Previous implementation used a boolean value.
  - Set to + if lexeme has all necessary inflection.
  - Set to − if lexeme needs more inflection.
  - Inadequate for some languages. Figure 1 shows an intended, ideal outcome, while Figure 2 shows inelegant results for a minimally different configuration.

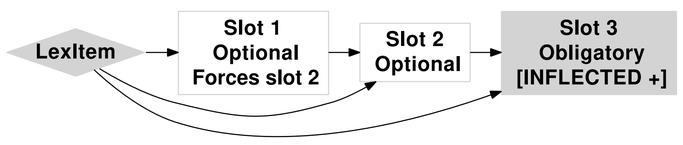

Figure 1: *Forcing before an obligatory slot.*

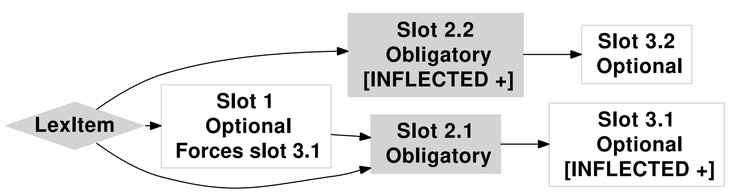

Figure 2: *Forcing around an obligatory slot.*

- New system makes INFLECTED take a complex value, and customizes features on that value for each grammar.
- Values inside INFLECTED generalized from *bool*-valued, to *luk* (Flickinger, 2000). See Figure 3.

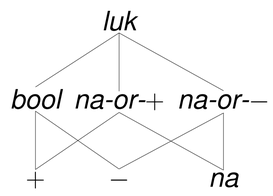

Figure 3: *The Luk value hierarchy*

+	morpheme has occurred
−	morpheme must and has not yet occurred
luk	initial condition for all flags, unless specifically further constrianed
na-or-+	satisfied condition for all flags

Lexical Rule Flags: Keeping track of morpheme occurrence.

- O'Hara (2008) defined a series of TRACK variables to keep track of the rules that have applied.
  - Separate from INFLECTED, and thus has no direct bearing on whether a lexeme can be used in a phrase.
- Our new system's **flags** are similar in principle, but are defined on *inflected*.
  - Directly affect lexeme's usability in phrases.
  - More nuanced than just a boolean value.
  - Allows for disjunction in rule occurrence requirements (see Figure 4).

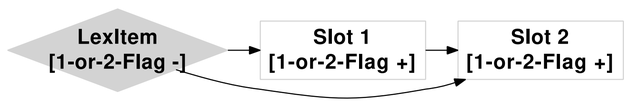

Figure 4: *Sequential disjunction.*

- Both lexical types and slots can change flag values.
  - Thus, the notion of obligatory slots is deprecated.
  - Rather, a lexical type requires a rule by affecting its flag (see Figure 5).

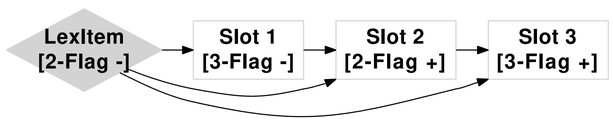

Figure 5: *Flag-based solution for configuration in Figure 2.*

3. An Example: Maltese

Object markers: Requirement dependent on lexical type.

- Object markers are obligatory with dropped objects and impossible with overt objects.

(1) Norma tikteb l-ittra
Norma ktb-ie-tvCCvC l-ittra
Norma write-3SG.FEM.IMPF DEF-letter
'Norma writes the letter.' [mlt]

(2) Norma tikteb-ha
Norma ktb-ie-tvCCvC-ha
Norma write-3SG.FEM.IMPF-3SG.FEM
'Norma writes it.' [mlt]

(3) *Norma tikteb-ha l-ittra.

(4) *Norma tikteb.

- Object markers don't attach to intransitive verbs.
- Object markers attach outside subject+aspect marking.
- The same subject+aspect markers are used for transitive and intransitive verbs.
- Desired analysis (Saleem, 2010):
  - All verbs have an obligatory subject+aspect marking slot.
  - Transitive verbs have an obligatory object marking slot, which includes a zero-marked "no droppping" morpheme.

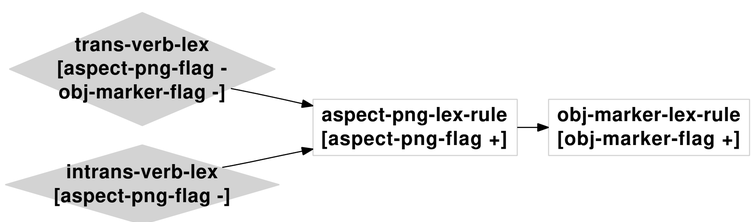

Figure 6: *Maltese lexical rules.*

HPSG types

Figure 7: *The* `inflected` *hierarchy.*

$$\left[\begin{array}{l} \textit{binary-phrase} \\ \text{ARGS} \left\langle \left[\begin{array}{ll}\text{INFLECTED} & \textit{infl-satisfied}\end{array}\right], \left[\begin{array}{ll}\text{INFLECTED} & \textit{infl-satisfied}\end{array}\right] \right\rangle \end{array}\right]$$

Figure 8: *Phrasal rules require lexemes to have satisfied inflectional flags.*

$$\left[\begin{array}{ll} \textit{inflected} & \\ \text{OBJ-MARKER-FLAG} & \textit{luk} \\ \text{ASPECT-PNG-FLAG} & \textit{luk} \end{array}\right]$$

Figure 9: *Flags defined on* `inflected`.

$$\left[\begin{array}{ll} \textit{infl-satisfied} & \\ \text{OBJ-MARKER-FLAG} & \textit{na-or-+} \\ \text{ASPECT-PNG-FLAG} & \textit{na-or-+} \end{array}\right]$$

Figure 10: *Satisfied condition set in* `infl-satisfied`.

$$\left[\begin{array}{ll} \textit{trans-verb-lex} & \\ \text{INFLECTED} & \left[\begin{array}{ll}\text{OBJ-MARKER-FLAG} & -\\ \text{ASPECT-PNG-FLAG} & -\end{array}\right] \end{array}\right]$$

Figure 11: *Trans. verbs require require two slots.*

$$\left[\begin{array}{ll} \textit{intrans-verb-lex} & \\ \text{ASPECT-PNG-FLAG} & - \end{array}\right]$$

Figure 12: *Intransitive verbs only require aspect markers.*

$$\left[\begin{array}{ll} \textit{obj-marker-lex-rule} & \\ \text{INFLECTED} & \left[\begin{array}{ll}\text{OBJ-MARKER-FLAG} & +\\ \text{ASPECT-PNG-FLAG} & \boxed{1}\end{array}\right] \\ \text{DTR.INFLECTED} & \left[\begin{array}{ll}\text{ASPECT-PNG-FLAG} & \boxed{1}\end{array}\right] \end{array}\right]$$

Figure 13: *Object marker lexical rule satisfies requirement.*

4. Evaluation

Regression Tests: Ensuring we don't lose coverage.

- Over 130 sample grammars and associated test suites covering many different languages and other (artificial) configurations are routinely checked for any loss in coverage.
  - Includes O'Hara's (2008) test languages, which were selected specifically for their morphological complexity.
- Create new regression tests for the specific cases the new system is intended to solve.
  - Requirements specified on lexical types.
  - Disjunctive requirements.

5. System Availability

`http://www.delph-in.net/matrix/customize/`

References

Emily M. Bender, Scott Drellishak, Antske Fokkens, Michael Wayne Goodman, Daniel P. Mills, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar prototyping and testing with the LinGO Grammar Matrix customization system. In *Proceedings of ACL 2010 Software Demonstrations*.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at COLING 2002*, pages 8–14.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1) (Special Issue on Efficient Processing with HPSG):15 – 28.

Kelly O'Hara. 2008. *A Morphotactic Infrastructure for a Grammar Customization System*. Master's thesis, University of Washington.

Safiyyah Saleem. 2010. *Argument Optionality: A New Library for the Grammar Matrix Customization System*. Master's thesis, University of Washington.

Gregory T. Stump. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge Univ Press.

Based on work supported by NSF grant #BCS-0644097