Morphological Marking of Constituent Questions. A Case for Nonlocal Amalgamation

Olga Zamaraeva University of Washington olzama@uw.edu

Abstract

This paper considers the role of NONLOCAL amalgamation in a cross-linguistic system of analyses which intend to support typologically diverse languages. NONLOCAL amalgamation (Bouma et al., 2001) was suggested in particular to get rid of extraction rules in Pollard and Sag's (1994) analysis of long-distance dependencies. However, in implemented projects like the English Resource Grammar (Flickinger, 2000, 2011) and the Grammar Matrix (Bender et al., 2002, 2010), the extraction rules have been maintained, while NONLOCAL amalgamation is used for the analysis of phenomena like the *easy*-adjectives. Zamaraeva and Emerson (2020) argue that, if extraction rules are kept, support of the English *easy*-adjectives may be an insufficient reason to maintain NONLOCAL amalgamation in a cross-linguistic system like the Grammar Matrix, as it complicates the analysis of multiple fronting. However, I present here a case of morphological marking of questions which further motivates NONLOCAL amalgamation, as the analysis is remarkably more simple with it than it is without it. An analysis of morphological marking of questions needs to be part of a cross-linguistic system like the Matrix as well as an analysis of multiple fronting, which adds a new tension at the level of the Matrix "core".

1 Morphological Marking of Constituent Questions

In some languages, questions (polar and constituent), are marked morphologically on the verb, as illustrated by examples from Negidal ([neg]; Tungusik), where the subject agreement marker on the verb in a declarative sentence (1) is different from that in an interrogative sentence (2)–(3).[1] Furthermore, in some languages, e.g. in Makah ([myh], Wakashan), the paradigms for constituent and polar questions are distinct. The marker occurring in constituent ('content') interrogatives is shown in (4)–(5); note that the question word here is actually analyzed as the main predicate (Davidson, 2002, p. 285). Polar question marking is exemplified in (6). The goal of this paper is to model separate paradigms as in Makah (4)–(6). Furthermore, the analysis is intended for the Grammar Matrix framework, which means it must fit into a broader system of analyses for different types of languages.

2 Background

2.1 The Grammar Matrix

The Grammar Matrix (Bender et al., 2002, 2010) is a meta-grammar engineering system. Given a typological specification (e.g. "the language uses special morphological marking in questions"), it automatically outputs an HPSG grammar fragment which covers data from the language to the extent specified. The grammar consists of lexical entries as well as phrasal and lexical rules. The types are customized according to the specification but each grammar is based on the same "core". For example, there is a core type for lexical rule, *lex-rule*, from which all customized lexical rules inherit. The project is implemented in DELPH-IN Joint Reference Formalism (Copestake, 2000).

[1] The future tense markers differ as well, and there is an additional question marker *=i*, but for the purposes of illustrating the phenomenon, it is sufficient to focus on just the person and number marker.

Morphological rules in the Matrix are lexical rules which apply to terminal nodes in the derivation (O'Hara, 2008; Goodman et al., 2015). For example, the Negidal example (2) would be analyzed in a Matrix-derived grammar as illustrated in (7). The affixes attach in order. I assume here that the final *=i* is a clause-final question particle which in the Grammar Matrix is analyzed as a complementizer (Bender and Flickinger, 2005). The tense as well as the person and number affixes, being specific to the interrogative paradigm, constrain the clause to have question semantics via the Sentential Force feature ([SF *ques*]), and all other affixes have to unify with that, so an affix from an indicative paradigm would not appear.

2.2 Nonlocal Amalgamation

The classic analysis of interrogatives in HPSG (Pollard and Sag, 1994; Ginzburg and Sag, 2000) relies on nonlocal features which introduce long-distance dependencies at the bottom of the derivation, specifically SLASH and QUE. SLASH introduces the dependency at the level of the verb while QUE does so at the level of the question word. Both features ultimately play a role in licensing question constructions: SLASH is instrumental in modeling fronting while QUE is instrumental in modeling "pied piping" (Ross, 1967). While motivated originally by data from English, the analysis is applicable cross-linguistically and is used in the Grammar Matrix (Zamaraeva and Emerson, 2020; Zamaraeva, forth.).

Bouma et al. 2001 suggested improvements to the analysis in Pollard and Sag 1994, which were later incorporated into Ginzburg and Sag's (2000) account of English interrogatives. The aspect of Bouma et al.'s (2001) analysis that I would like to consider here is the NONLOCAL amalgamation principle (originally called SLASH amalgamation). The NONLOCAL amalgamation principle constrains a word's NONLOCAL features to be the union of its arguments' NONLOCAL sets, which then allows phrases to simply inherit the NONLOCAL values of the head daughter, instead of explicitly gathering all the values of all daughters. One of the main motivations for NONLOCAL amalgamation was getting rid of extraction rules. In Pollard and Sag 1994, fronted subjects, complements, and adjuncts are extracted explicitly by the corresponding lexical rules. In Bouma et al. 2001 and subsequently in Ginzburg and Sag 2000, extraction rules were obviated by directly positing lexical entries with *gap*s in the place of arguments.

In the English Resource Grammar (Flickinger, 2000, 2011) and subsequently in the Grammar Matrix (Bender et al., 2002, 2010), NONLOCAL amalgamation was implemented, but mainly to support an elegant analysis of e.g. English *easy*-adjectives. Extraction rules were maintained, in particular to avoid positing additional lexical entries for all heads which can have their arguments extracted. Zamaraeva and Emerson (2020) argue that while for the English Resource Grammar implementing NONLOCAL amalgamation may be convenient, maintaining NONLOCAL amalgamation in the Grammar Matrix complicates the analysis of multiple question fronting such as is found in Russian. They show that, in the context where extraction rules are present anyway, maintaining NONLOCAL amalgamation necessitates adding even more extraction rules, to model languages with flexible order of fronting, where an extracted adjunct may appear between two extracted arguments as well as before or after them. In §3, I present a counterpoint to this and show that NONLOCAL amalgamation greatly simplifies the analysis of morphological marking of questions.

3 Analysis

In the context of question marking, a lexical rule can be (i) indicative or (ii) interrogative; and if it is interrogative, it can serve (a) polar, (b) constituent questions, or (c) both. Morphological marking of constituent questions is an example of where NONLOCAL amalgamation allows for a particularly elegant analysis, especially when it comes to modeling the distinction between options (a), (b), and (c) above, while modeling the difference between (i) and (ii) is straightforward with or without NONLOCAL amalgamation.

Consider a hierarchy of lexical rules (8). An *indicative-lex-rule* simply says its SF value is *prop*. This ensures the correct semantics. An *interrogative-lex-rule*, on the other hand, constrains its SF value to be

ques, allowing verbs marked with such rule to make the semantics of the clause interrogative without an additional phrase structure rule. In (7), there is also a question particle but in some languages there will not be one and an interrogative clause will be licensed simply via the head-subject rule or an argument drop rule. The distinction between (i) and (ii), and by extension (c),[2] is straightforwardly modeled with just the SF feature. An affix which is [SF *ques*] cannot appear in the same sequence of lexical rules as an affix which is [SF *prop*] because the SF value is identified between the mother and the daughter in such sequences. This is true with or without NONLOCAL amalgamation. Modeling the distinction between (a) and (b), however, is much easier with NONLOCAL amalgamation than without it.

Under the NONLOCAL amalgamation assumption, a verb's QUE value will be the append of its subject's and object's (9). In other words, if one or more of the verb's arguments are *wh*-words, the verb's own QUE list will be non-empty; otherwise it will be empty. Given this, modeling the distinction between (i) and (ii) and furthermore between (a) and (b) is straightforward. Markers which are to be used exclusively in polar questions constrain the daughter of the rule (the verb) to be QUE-empty (10). Under the NONLOCAL amalgamation assumption, this is the same as to say that neither of the arguments is a *wh*-word. Conversely, the ones which are to be used exclusively for *wh*-questions are customized to take QUE-nonempty daughters (11).[3] This means one or more of the arguments is a *wh*-word. I illustrate the situation with subtrees for (6) and (4) presented side by side in (12). In the analysis on the left, because the verb has no *wh*-argument, it has an empty QUE value. In the analysis on the right, the verb has a *wh*-subject, and so its own QUE value is not empty. The correct morphological behavior follows, with only the appropriate affix licensed in each case. Note that there is no need to worry about how many arguments the verb has and how many of them are *wh*, and which positions they occupy on the argument list. If the same marker is used for both polar and constituent questions (c), the QUE value on the daughter is underspecified.

Without NONLOCAL amalgamation, option (c) (languages which just use one marker for all types of questions, like Negidal) still does not pose complications; the analysis is the same as with NONLOCAL amalgamation as the QUE value of the rule's daughter is underspecified on the rule. However, neither (10) nor (11) will lead to the desired analysis of (a) and (b). Lexical entries which are not *wh*-words have empty QUE-lists, and without NONLOCAL amalgamation, the way the QUE values are propagated up the tree is an explicit inheritance in unary and an explicit append in binary rules. Verbs do not take their arguments' QUE values and their own QUE lists are empty. This means that most verbs will unify with the daughter of (10) regardless of what their arguments are. An explicit constraint must be put on the verb's arguments instead, but, in DELPH-IN JRF, this means an explicit constraint is required for SUBJ and yet another for COMPS. The *interrogative-lex-rule* should now have at least three subtypes, *polar*, *wh-subj*, and *wh-obj*, and the latter two should be furthermore constrained to not both apply where one is sufficient.

4 Conclusion

NONLOCAL amalgamation (Bouma et al., 2001) definitely makes sense to use in an analysis of English (Flickinger, 2000, 2011); perhaps it makes less sense to use it in an analysis of Russian, so long as it is an analysis based on extraction rules (Zamaraeva and Emerson, 2020). I show in this paper that there are phenomena beyond the English *easy*-languages for which NONLOCAL amalgamation greatly simplifies the analysis. The question then is, what should the choice be in a cross-linguistic system like the Grammar Matrix (Bender et al., 2002, 2010). The tension here is between adding more types to the type hierarchy (e.g. *wh-subj/obj-lex-rule* and a parametrized list to rule out spurious ambiguity) and having multiple adjunct extraction rules accommodating the various word orders, as explained in Zamaraeva and Emerson 2020.

[2] Option (c) is essentially a statement that only the distinction between (i) and (ii) is relevant.

[3] *Cons* here is a type for non-empty list.

There is an additional, practical dimension, that of the ease of reasoning about a complex system such as the Grammar Matrix; Zamaraeva and Emerson (2020) note that in practice, NONLOCAL amalgamation makes the system more error-prone. I do not offer an answer to the question: Should NONLOCAL amalgamation be part of the Grammar Matrix. But modeling multiple question fronting like in Russian and morphological question marking like in Makah using the same Matrix core presents a challenge which ultimately relates to the discussion of whether all languages should be using the same core at all (Haspelmath, 2010).

5 Data and Figures

(1) oǯa-va iche-ǯee-**v**
track-ACC see-FUT-**1SG**
'I will see the tracks.' [neg] (Hölzl, 2018, p. 295)

(2) ii-ǰə-**m**=i?
enter-FUT.Q-**1SG.Q**=Q
'Shall I come in?' [neg] (Hölzl, 2018, p. 295)

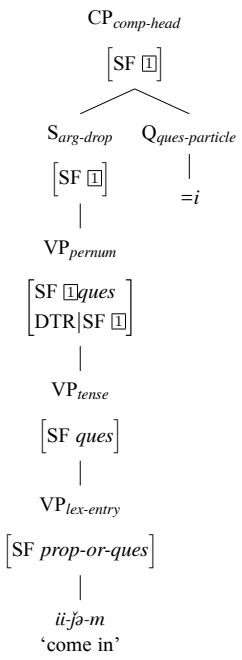
(3) eeva iche-ǯa-**m**?
what see-FUT.Q-**1SG.Q**
'What will I see?' (Hölzl, 2018, p. 295)

(4) ʔačaq=qaːɬ dudu$^{\text{l}}$k
who=CONTENT.3SG sing
'Who is singing?' [myh] (Davidson, 2002, p.285)

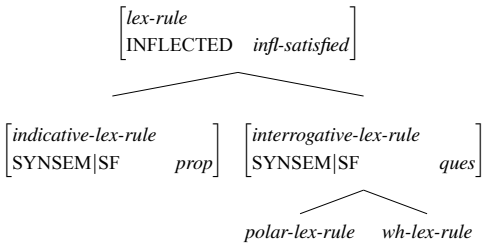
(5) baqiq=qaːɬ ti$^{\text{l}}$
what=CONTENT.3SG DEM
'What is this?' [myh] (Davidson, 2002, p.285)

(6) dudu$^{\text{l}}$k='aƛ̓=qaːk=s
sing=TEMP=POLAR=1SG
'Am I singing?' [myh] (Davidson, 2002, p.100)

(7)



(8)

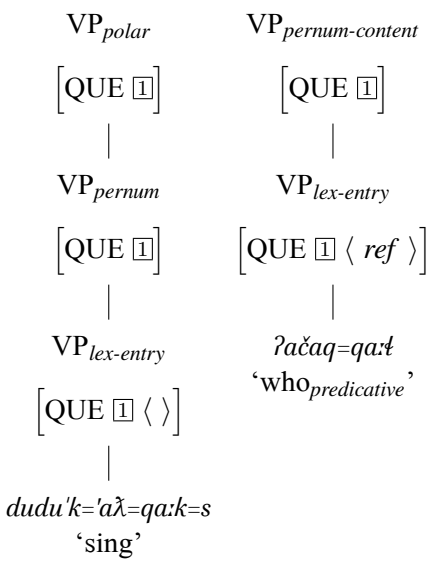


(9)
$$\begin{bmatrix} \textit{verb} & \\ \text{SYNSEM} & \begin{bmatrix} \text{NON-LOCAL|QUE|APPEND} & \langle \boxed{1}, \boxed{2} \rangle \\ \text{LOCAL|CAT|VAL} & \begin{bmatrix} \text{SUBJ} & \left\langle \left[\text{NON-LOCAL|QUE } \boxed{1} \right] \right\rangle \\ \text{COMPS} & \left\langle \left[\text{NON-LOCAL|QUE } \boxed{2} \right] \right\rangle \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

(10)
$$\begin{bmatrix} \textit{polar-lex-rule} & \\ \text{SYNSEM}|\text{SF} & \textit{ques} \\ \text{DTR}|\text{SYNSEM}|\text{NON-LOCAL}|\text{QUE} & \langle\,\rangle \end{bmatrix}$$

(11)
$$\begin{bmatrix} \textit{wh-lex-rule} & \\ \text{SYNSEM}|\text{SF} & \textit{ques} \\ \text{DTR}|\text{SYNSEM}|\text{NON-LOCAL}|\text{QUE} & \textit{cons} \end{bmatrix}$$

(12)



References

Emily M. Bender and Dan Flickinger. 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 (Posters/Demos)*, Jeju Island, Korea, 2005.

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 John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan, 2002.

Emily M. Bender, Scott Drellishak, Antske Sibelle Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72. ISSN 1570-7075.

Gosse Bouma, Robert Malouf, and Ivan A Sag. 2001. Satisfying constraints on extraction and adjunction. *Natural Language & Linguistic Theory*, 19(1):1–65.

Ann Copestake. 2000. Appendix: Definitions of typed feature structures. *Natural Language Engineering*, 6(01):109–112.

Matthew Davidson. 2002. *Studies in Southern Wakashan (Nootkan) grammar*. PhD thesis, University of New York at Buffalo, Buffalo, NY.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(01):15–28.

Dan Flickinger. 2011. Accuracy v. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.

Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

Michael Wayne Goodman, Joshua Crowgey, Fei Xia, and Emily M. Bender. 2015. Xigt: Extensible interlinear gloss text for natural language processing. *Language Resources and Evaluation*, 49 (2):455–485.

Martin Haspelmath. 2010. Framework-free grammatical theory. In *The Oxford handbook of linguistic analysis*, pages 341–365. Oxford University Press.

Andreas Hölzl. 2018. *A typology of questions in Northeast Asia and beyond: An ecological perspective*. Language Science Press, Berlin.

Kelly O'Hara. 2008. *A morphotactic infrastructure for a grammar customization system*. PhD thesis, University of Washington.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

John Robert Ross. 1967. *Constraints on variables in syntax*. PhD thesis, Massachusetts Institute of Technology.

Olga Zamaraeva. forth. *A cross-linguistic analysis of constituent questions for the Grammar Matrix*. PhD thesis, University of Washington.

Olga Zamaraeva and Guy Emerson. 2020. Multiple question fronting without relational constraints: An analysis of Russian as a basis for cross-linguistic modeling. In *Proceedings of the 27th International Conference on Head-Driven Phrase Structure Grammar*, page to appear, Virtual conference, 2020.