Multiple Question Fronting without Relational Constraints. An Analysis of Russian as a Basis for Cross-Linguistic Modeling

Olga Zamaraeva
University of Washington
olzama@uw.edu

Guy Emerson
University of Cambridge
gete2@cam.ac.uk

Abstract

We describe an account of multiple question fronting in HPSG that assumes only unification as a basic operation. We develop and test our analysis in the context of data from Russian, and we integrate parts which can be used cross-linguistically into the Grammar Matrix, a grammar engineering toolkit. To do so we revise the Grammar Matrix core, which previously disallowed multiple question fronting and relied on *difference lists* for appending non-local lists. We utilize the recently proposed *append list*, which allows us to manipulate lists more easily. Additionally, to rule out spurious ambiguity in the order of application of rules, we posit a new list type which cannot contain gap elements.

1 Fronting of Question Phrases

Here we take prototypical constituent (aka *wh-*) questions to be a conventional and direct way of asking for information (Idiatov, 2007):

(1) Who arrived? [eng]

(2) Who saw what? [eng]

(3) Who do you think arrived? [eng]

(4) Kto chto videl?
who.NOM what.ACC see.PST.3SG
'Who saw what?' [rus]

Constituent questions are a case of *unbounded dependency constructions* (UDCs; Gazdar, 1981) meaning that the question phrase appears outside of the boundary of the clause to which it belongs (3). Languages differ with respect to how many question phrases can front. Famously in Slavic languages, all question phrases may be fronted (4). On the surface, fronting is obligatory in some languages (Bulgarian) and optional in others (Russian). We second Przepiórkowski (1998) in noting that the discussion on whether fronting is obligatory or optional in Slavic amounts to the discussion of superiority effects (Chomsky, 1973), which in HPSG as well as LFG has not been treated as a purely syntactic phenomenon (Ginzburg and Sag, 2000; Mycock, 2006). In particular, examples like (5a)-(5b) are perfectly acceptable in contexts where there is a salient emphasis (arising from the discourse) distributed between both arguments of the verb (cf. Pesetsky, 1987). Such sentences should be analyzed taking information structure into account.

(5) a. Chto kto videl?
what.ACC who.NOM see.PST.3SG
'Who saw what?' [rus]

b. Kto videl chto?
who.NOM see.PST.3SG what.ACC
'Who saw what?' [rus]

Finally, Russian does not exhibit obligatory pied piping (Ross, 1967), so both (6a) and (6b) are grammatical:

(6) a. Kakuju knigu Ivan chitaet?
which.ACC book.ACC Ivan.NOM read.3SG
'Which book is Ivan reading?' [rus]

b. Kakuju Ivan chitaet knigu?
which.ACC Ivan.NOM read.3SG book.ACC
'Which book is Ivan reading?' [rus]

Our goal is to systematically account for the data where question words, one or more, may appear at the left edge of the clause. We do so by developing a Russian grammar fragment and integrating it into a cross-linguistic grammar engineering framework, the Grammar Matrix (§2; Bender et al., 2002, 2010). Integrating an analysis of question fronting into the system means extending (i) the range of hypotheses which can be rigorously tested with the system, such as various combinations of single/multiple and optional/obligatory fronting; and (ii) the platform itself so that new syntactic phenomena (such as relative clauses) can be modeled on top of and in interaction with our analysis.

2 Background: DELPH-IN Formalism and the Grammar Matrix

Our analysis is part of a larger project, a constituent questions library for the Grammar Matrix (Bender et al., 2002, 2010). As such, it is couched within the DELPH-IN joint reference formalism (JRF; Copestake, 2000), a restrictive variant of HPSG developed to balance expressivity with computational efficiency. It does not allow relational constraints such as shuffle operators (Reape, 1994) and furthermore requires that the number and order of daughters of each phrase structure rule be fixed in the definition of the rule, precluding systems that separate immediate dominance from linear precedence (e.g. Engelkamp et al., 1992; Kathol, 1995). Thus, our analysis builds on Pollard and Sag 1994 and Ginzburg and Sag 2000 and exists in parallel with linearization-based accounts of multiple fronting such as Penn 1998.

The goal of the larger project is to add cross-linguistic support for constituent questions to the Grammar Matrix system. The system includes a questionnaire that elicits typological and lexical information about a language from a linguist-user and a back-end logic that customizes the Matrix core grammar types according to the elicited specifications. The resulting grammar fragments are suitable for both parsing and generation and map between surface strings and Minimal Recursion Semantics (Copestake et al., 2005) representations. In addition to facilitating the development of grammars for practical applications, the system also can be used in linguistic hypothesis testing (Bender et al., 2008; Fokkens, 2014). We present an analysis of multiple question fronting which represents a hypothesis that data such as Russian can be accounted for with multiple application of the *filler-gap* rule, without relational constraints.

3 Related Work

Several grammars of Slavic languages written in DELPH-IN JRF exist (Avgustinova and Zhang, 2009; Osenova, 2010; Fokkens and Avgustinova, 2013) but none of them cover multiple questions. We build our analysis on the general treatment of UDCs in DELPH-IN JRF which owes largely to the English Resource Grammar (Flickinger, 2000, 2011). The ERG provides a basic type hierarchy for clauses, including *interrogative-clause* and *filler-gap-phrase*. These types, in turn, stem from foundational theoretical literature in HPSG. The *filler-gap* rule and the feature SLASH come from Pollard and Sag 1994. The hierarchy of clauses including *interrogative-clause*, the unary extraction rules, and the feature QUE which introduces interrogative semantics on the lexical level, come from Ginzburg and Sag 2000. Finally, *lexical threading* (Bouma et al., 2001) means verbs amalgamate their arguments' QUE and SLASH values, for a streamlined analysis of lexical entries which select for complements containing a gap, such as the English *easy*-adjectives. The treatment of UDCs differs between the aforementioned theoretical accounts and DELPH-IN, in that the ERG and subsequently most DELPH-IN grammars including the Matrix core have been restricted to only allow one extraction and therefore at most one fronted element (7)-(8).[1]

(7)
$$\begin{bmatrix} \textit{OLD-filler-gap-phrase} & \\ \text{SLASH} & \langle\rangle \\ \text{ARGS} & \left\langle \boxed{1}, \begin{bmatrix}\text{SLASH} \left\langle \boxed{1} \right\rangle \end{bmatrix} \right\rangle \end{bmatrix}$$

(8)
$$\begin{bmatrix} \textit{OLD-nonlocal} & \\ \text{SLASH} & \textit{0-1-dlist} \end{bmatrix}$$

These restrictions were placed in order to avoid unbounded recursive extraction, particularly of adjuncts. That the restriction had to be global is due to the fact that in DELPH-IN, list appends have so far been implemented using *difference lists* (for an exposition, see Csenki, 2010). One limitation of difference lists is that once a list has been stated to have a specific length, it can no longer be easily appended to another list. This limitation did not present an issue for the ERG because at most one extraction is in fact desired in English,[2] but it is a problem for the Grammar Matrix core which is meant to be cross-linguistically applicable. At the same time, Sag et al. (2003) offer an analysis of English with multiple extraction, a version of which, to our knowledge, was implemented in DELPH-IN JRF once before, by Crysmann (2015). While our analysis is similar to Crysmann 2015 in that it allows multiple values on the SLASH list, we also make it available for automated implementation via the Grammar Matrix system after testing it for general cross-linguistic applicability. This makes it possible to obtain grammars that license multiple fronting automatically, given a typological description. We also test the analysis on a medium-size Russian test suite. *Append-list*, a type which we use instead of difference lists and which does allow us to easily operate on lists of different specific lengths and types, was first suggested by Emerson (2017, 2019) and was first used by Aguila-Multner and Crysmann (2018) for gender resolution.[3]

4 Analysis

We revise the single extraction restriction in the Grammar Matrix which has been preventing it from scaling cross-linguistically in the dimension of UDCs. We test that our revisions do not negatively affect any of the existing analyses which rely on the Grammar Matrix core and customization logic.[4] First we fully replace all use of difference lists in the Grammar Matrix with append lists. Then, following Sag et al. (2003), we simply say that there may be more than one item on the SLASH list, and that after *filler-gap* applies and the first item on the list is discharged (a gap is filled), the rest of the SLASH list is passed up to the mother node (9). This discharging of an element on the SLASH list requires closing the end of the list, or else the grammar can always "hallucinate" another element. With difference lists, closing the list makes further appends impossible. However, with append lists, we avoid this problem: hallucination is blocked because lists are always closed, but further list appends are still possible.

(9)
$$\begin{bmatrix} \textit{NEW-filler-gap-phrase} & \\ \text{SLASH} & \boxed{2} \\ \text{ARGS} & \left\langle \boxed{1}, \begin{bmatrix}\text{SLASH} \left\langle \boxed{1} \right\rangle \oplus \boxed{2} \end{bmatrix} \right\rangle \end{bmatrix}$$

(10)
$$\begin{bmatrix} \textit{NEW-nonlocal} & \\ \text{SLASH} & \textit{append-list} \end{bmatrix}$$

[1]The AVMs are abbreviated, with only features relevant to the discussion shown.

[2]Multiple extraction may be required to account for the *violin-sonata* sentences; see Sag et al. 2003 for a detailed discussion.

[3]For technical details regarding append lists, refer to Emerson 2019.

[4]These analyses include specifications and tests suites from 417 artificial and 56 real languages, with corresponding "gold" semantic representations for each. They were added by multiple Grammar Matrix contributors over the years as support for more and more syntactic phenomena was added.

We change constraints such as in (7) to something similar to (9) throughout the Grammar Matrix core which serves also as the core of our grammar of Russian. This includes types for argument extraction and also any types which used to insist on SLASH-empty head-daughters and generally QUE-empty daughters, such as the *head-adjunct-phrase*. After introducing these and related changes which we will elaborate on in the full version of the paper, the grammar can account for (4)-(6a) and for a range of additional Russian data. Accounting for (6b) required an additional determiner extraction rule.

Revising the phrase structure rules such as *head-adjunct* to allow SLASH-ed head daughters requires us to address two problems: recursive adjunct extraction and spurious ambiguity. We note that both problems can be solved by positing new typed lists. In this abstract, we present the solution for spurious ambiguity.[5]

To allow multiple fronting, we allow the *head-adjunct* rule to have SLASH-ed head daughters. This however means the rule can apply both before and after e.g. subject extraction. We propose a simple extension to the Grammar Matrix core type hierarchy to solve this problem. We constrain the head daughter of the *head-adjunct-phrase* to not allow subjects of type *gap* while still allowing both canonical subjects and phrases where subject has already been saturated (and so the subject list is empty). This is easy to do given that originally, the type *gap* was specifically defined as non-unifiable with *canonical-synsem*. We posit a subtype of *list* which does not allow items of type *gap* on it (11) and then use that type as a constraint on the head daughter's subject (12).

(11) $\begin{bmatrix} \textit{canonical-nonempty-list} & \\ \text{FIRST} & \textit{canonical-synsem} \\ \text{REST} & \textit{canonical-list} \end{bmatrix}$ (12) $\begin{bmatrix} \textit{head-adjunct-phrase} & \\ \text{HEAD-DTR.SUBJ} & \textit{canonical-list} \end{bmatrix}$

We integrate our analysis into the Grammar Matrix system and add 11 new artificial language descriptions and test suites, featuring various combinations of the type (single or multiple) and the obligatoriness of fronting, to the Matrix regression test system. We also contribute a 375-sentence test suite from Russian that contains a variety of constituent questions as well as a number of declarative sentences intended to test that the basic treatment of word order variation is working properly. Our Russian grammar has 83% coverage and 7% overgeneration on this test suite. The unanalyzed grammatical sentences include coordinated questions of the type *Who and what...* for which we do not yet propose an analysis; the analyzed sentences which are marked ungrammatical in the test suite include partitive case constructions, the treatment of which would be orthogonal to the fronting analysis, but which we kept in the test suite for perspective and to better define future directions.

5 Conclusion

Following Crysmann (2015), we provide evidence that modeling multiple UDCs is possible with no relational constraints (as a mechanism separate from unification). We developed an analysis of Russian multiple question fronting and integrated it into the Grammar Matrix system. We achieve this using two types of lists previously not in the Grammar Matrix hierarchy: append lists (Emerson, 2019) and a new type (*canonical-list*) which is incompatible with *gap*.

[5]Bouma et al. (2001) argued for a unified account of argument and adjunct extraction; Przepiórkowski (1998) is of the opinion that the distinction between arguments and adjuncts does not play a consistent role in Slavic syntax. In principle, it is possible to argue that recursive adjunct extraction should be allowed and is only subject to processing constraints. Still, we intend to limit our grammar of Russian to single adjunct extraction. The solution would be very similar to the one described for ambiguity: We would posit a typed list for adjunct gaps and would restrict the adjunct-extraction rule's head daughter to have a SLASH list of non-adjunct gaps. We will include this in the full version of the paper.

References

Gabriel Aguila-Multner and Berthold Crysmann. 2018. Feature resolution by lists: The case of French coordination. In Annie Foret, Greg Kobele, and Sylvain Pogodalla, editors, *Formal Grammar. 23rd International Conference*, Lecture Notes in Computer Science, pages 1–15. Springer, Heidelberg.

Tania Avgustinova and Yi Zhang. 2009. Exploiting the Russian National Corpus in the development of a Russian Resource Grammar. In *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, pages 1–11. Association for Computational Linguistics.

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, Dan Flickinger, and Stephan Oepen. 2008. Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X conference: Computational linguistics for less-studied languages*, pages 16–36. Citeseer.

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.

Noam Chomsky. 1973. Conditions on transformations. In Anderson and Kiparsky, editors, *A Festschrift for Morris Halle*, pages 232–286. Holt, Rinehart & Winston, NY.

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

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2-3):281–332.

Berthold Crysmann. 2015. Resumption and extraction in an implemented HPSG grammar of Hausa. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 65–72.

Attila Csenki. 2010. Difference lists in prolog.

Guy Emerson. 2017. (Diff)list appends in TDL. URL `http://www.delph-in.net/2017/append.pdf`. DELPH-IN summit, Oslo, Norway.

Guy Emerson. 2019. Wrapper types: relational constraints without relational constraints. URL `http://users.sussex.ac.uk/ johnca/summit-2019/wrapper-types.pdf`. DELPH-IN summit, Cambridge, UK.

Judith Engelkamp, Gregor Erbach, and Hans Uszkoreit. 1992. Handling linear precedence constraints by unification. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pages 201–208. ACL.

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.

Antske Sibelle Fokkens. 2014. *Enhancing Empirical Research for Linguistically Motivated Precision Grammars*. PhD thesis, Department of Computational Linguistics, Universität des Saarlandes.

Antske Sibelle Fokkens and Tania Avgustinova. 2013. SlaviCLIMB: Combining expertise for Slavic grammar development using a metagrammar. In *Workshop on high-level methodologies for grammar engineering*, pages 87–92.

Gerald Gazdar. 1981. Unbounded dependencies and coordinate structure. In *The Formal Complexity of Natural Language*, pages 183–226. Springer.

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

Dmitry Idiatov. 2007. *A typology of non-selective interrogative pronominals*. PhD thesis, University of Antwerp.

Andreas Kathol. 1995. *Linearization-based German syntax*. PhD thesis, The Ohio State University.

Louise Mycock. 2006. *The Typology of Constituent Questions: A Lexical-Functional Grammar analysis of Wh-questions*. PhD thesis, University of Manchester.

Petya Osenova. Bulgarian resource grammar–efficient and realistic. Technical report, Stanford University, CSLI, 2010.

Gerald Penn. 1998. Linearization and Wh-extraction in HPSG: Evidence from Serbo-Croatian. In *Slavic in HPSG*.

David Pesetsky. 1987. Wh-in-situ: Movement and unselective binding. *The representation of (in) definiteness*, 98:98–129.

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.

Adam Przepiórkowski. 1998. On complements and adjuncts in Polish. In *Slavic in HPSG*.

Mike Reape. 1994. *A formal theory of word order: A case study in West Germanic*. PhD thesis, University of Edinburgh.

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

Ivan A Sag, Thomas Wasow, Emily M. Bender, and Ivan A Sag. 2003. *Syntactic theory: A formal introduction*. Center for the Study of Language and Information Stanford, CA.