Shared and Non-shared Grammar in Modelling Slavic Morphosyntax

Tania Avgustinova and **Hans Uszkoreit**

DFKI Language Technology Lab
German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
{avgustinova,Uszkoreit}@dfki.de

Abstract

In line with the increasing interest in fine-grained modular linguistic representations, we investigate the plausibility of exploiting shared and non-shared portions of grammars for the systematic and efficient development of grammatical resources for multiple languages. We use a new systematic structuring of well-known phenomena form Slavic morphosyntax to illustrate our approach to the design of shared grammars.

1. Introduction

There is an emerging awareness in the grammar engineering community that for the rapid development of grammars for new languages and for the systematic adaptation of grammars to variants of languages, the notion of grammar sharing is essential, as it aims at maximum reusability, lifting out the elements that can and should be common across language specific grammars. Initial efforts in designing an open-source starter-kit for rapid development of computational grammars in terms of a grammar matrix with associated grammar engineering environment have been reported by (Bender, Flickinger & Oepen, 2002). Emerging in parallel to the grammar sharing concept of (Avgustinova, 2002), this early version of the grammar matrix comprises first and foremost the formalism-specific technical devices and the basic feature geometry, while on the linguistic side, only general semantic types in the formalisation of (Copestake, Lascarides & Flickinger, 2001) are envisaged. In line with increasing interest in fine-grained modular linguistic representations, it is essential to investigate the plausibility of designing linguistically motivated shared grammatical resources that would systematically go beyond language-independent and fairly abstract semantic information.

However, whereas the grammar matrix approach defines a pragmatically selected, nicely categorized and steadily growing collection of possible building blocks for individual grammars, we aim at a systematic structuring of the space of potential phenomena so that each possible grammatical relationship can be composed of atomic elements.

In this sense, our approach to the specification of shared grammar continues a program first proposed by (Kameyama, 1988). Megumi Kameyama advocates a fine-grained "atomized" structure of grammatical knowledge representation and proposes the utilization of lattice-structured inheritance graphs for the composition of shared knowledge. As the smallest units, she uses feature structure templates, i.e. AVM macros that are ordered in a semi-lattice. This makes it hard to define a clean semantics for units and their interaction.

Today, we can exploit finely grained type hierarchies as they are used in HPSG for the systematic multidimensional structuring of the wide inventory of potential grammar ingredients thus allowing components of actual grammars to inherit the appropriate combinations of phenomena. The semantics of the typed-feature-structure formalism is powerful

enough to formally describe the effects of combining atomic building blocks.

2. The Slavic Connection

As Slavic languages share a wider range of grammatical properties than typically reflected in standard multilingual applications, a theoretically motivated and linguistically sound modularity might certainly incorporate important insights from Slavic linguistics.

According to traditional morphosyntactic descriptions of the Slavic language family, Southeast Slavic, and more specifically Bulgarian, represents an extreme as a language lacking morphological cases and infinitive but showing an impressively complex verbal system, a definite article, a full-fledged clitic paradigm, and the phenomenon of clitic doubling.

Another extreme is claimed by Russian, which is a prototypical East Slavic language having morphological case and infinitive but lacking any auxiliary or pronominal clitics and extensively employing non-verbal predication. Therefore, careful consideration of Bulgarian and Russian data is indispensable for the purposes of the current study.

A selective focusing on key phenomena from West Slavic languages, represented by Czech and Polish, is crucial too. Together with Southeast Slavic, represented by Serbo-Croatian and Slovene, they have clausal-domain auxiliary and pronominal clitics, along with morphological case and infinitive.

The *common properties* of Slavic languages have been observed – both in literature and related research – at various intermediate levels of linguistic abstraction. The interesting question arises whether *minimal differences* are also detectable as parameters of systematic variation. This has been the starting point of our own investigation.

3. The DELPH-IN Initiative

DELPH-IN[1] as a collaborative effort aims at deep linguistic processing of human language, with the goal of combining linguistic and statistical processing methods for getting at the meaning of texts and utterances. There is a shared commitment to re-usable, multi-purpose resources and active exchange. Based on contributions from several members and joint development over many years, an open-source repository of software and linguistic resources has been created that already enjoys wide usage in education, research, and application building.

As HPSG[2] implementations evolved for several languages within the same common formalism, it became clear that homogeneity among existing grammars could be increased and development cost for new grammars greatly reduced by compiling an inventory of cross-linguistically valid (or at least useful) types and constructions. The LinGO[3] Grammar Matrix[4] provides a starter kit to grammar engineers, which (in its current release version 0.4) comprises (a) types definitions for the basic feature geometry and technical devices, (b) the representation and composition machinery for Minimal Recursion Semantics in a type feature structure grammar, (c) general classes of rules, including derivational and inflectional (lexical) rules, unary and binary phrase structure rules, headed and non-headed rules, and head-initial and head-final rules, and (d) types for basic constructions such as head-complement, head-specifier, head-subject, head-filler, and head-modifier rules, coordination, as well as more specialized classes of constructions. Work in progress concentrates on an initial set of modules to handle recurring yet non-universal patterns, including word order systems, different strategies for sentential negation, and different strategies for yes-no questions.

[1] http://www.delph-in.net/
[2] http://hpsg.stanford.edu/
[3] http://lingo.stanford.edu/
[4] http://www.delph-in.net/matrix/

The goals of the project in the long term are to create a tool that allows field linguists to easily build implemented grammars as they research a language, to test hypotheses and encode their results; and to facilitate the exchange of data and analyses of a wide range of phenomena across diverse languages.

Thus in the current setup, the matrix is a box filled with a growing number of building blocks for grammars. The building blocks have been designed as part of the development of an existing HPSG grammar.

As will become apparent in the following sections, our approach is different. We are trying to chart the range of potential phenomena, so that many of them are already covered, even though we have no grammar for them yet. The structure of the matrix could to some degree reflect the family relations among human languages. In this sense, the matrix would not just be a reservoir but also a tool for determining the location of a language in a many-dimensional typological space. It should also be suited for modelling the minimal steps in language change or dialectal variation.

Nevertheless, we believe that our approach is compatible with the current matrix program. It is quite likely that not all phenomena classes and language families can be captured in our systematic way or at least that this endeavor would take a very long time. Therefore, a mixed approach somewhere between theoretical ambition and practical concession may be the best solution.

4. Shared Grammar for Slavic

The grammar sharing concept of (Avgustinova, 2002) highlights a somewhat different perspective: a common core module which combines with a number of extensions designed to be consistent with the core. For a language family, the core module is expected to be relatively large and to cover the major phenomena areas. It is also abstract enough in order to be shared by all Slavic languages modulo the appropriate further specification. Intuitively, the core incorporates what is interpretable as "typical Slavic". The extensions can be of different granularity in order to encode properties and phenomena that are characteristic of respective sub-groups, but need not be attested in other members of the family. Yet, all these phenomena would constitute natural extensions of the common core module. The key question is how to achieve such grammar architecture, if it is only obvious that whatever one chooses to dub "common" or "shared" need not be available – or at least not to equal extent – in every specific Slavic language or dialect. An appropriate level of ontological representation is required. Thus, systematic relations motivate in (Avgustinova and Uszkoreit, 2000) shared patterns of variation cross-linguistically and across constructions. A related meta-grammar perspective is taken in (Avgustinova, 2003) to distinguish languages with rich morphology from those with impoverished morphology in a principled way, and yet, to directly express linguistic generalizations of various degrees of abstraction. The ontological level can be encoded as compatible multidimensional hierarchies of types of linguistic entities, with constraint inheritance from more general to more specific types.

With regard to formalization, a class of constraints called relational dependencies provides a universal means of introducing more abstract and modular specifications in grammar and lexicon (Dörre et al., 1992). Relational dependencies are constraints that hold among typed feature structures. If we allow relational dependencies as part of our grammar specification language, they can be used within the specified types. They are constraints on permissible values of features with respect to other values. Since we have based our notion of grammatical relationships on binary dependencies, we only need binary relational dependencies. Relational dependencies themselves can be expressed as feature structures with two attributes. These feature structures themselves can be typed. The types can be ordered in a multiple-inheritance

hierarchy, preferably a semi-lattice. In this way we can construct a formal specification of the hierarchy of dependencies.

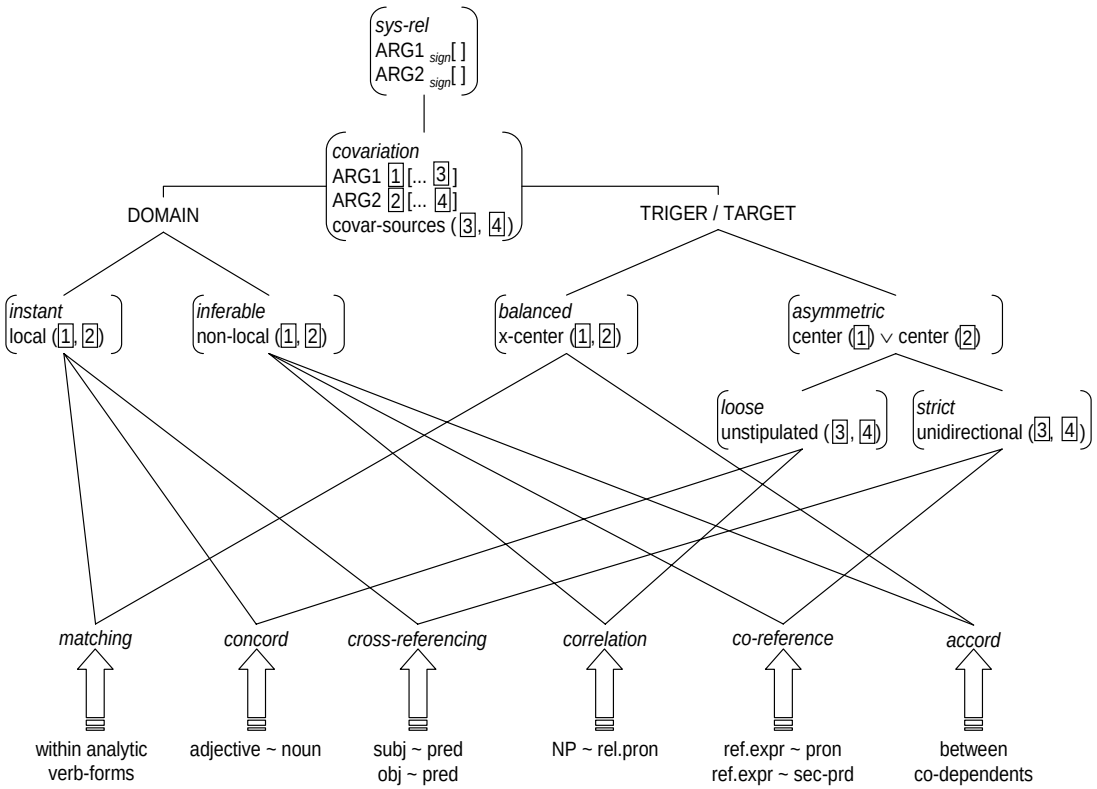
Let us now look at the broad spectrum of agreement phenomena that constitutes a challenge to any linguistic theory maintaining a universality claim and to any theoretically grounded typological description. Syntagmatic regularities in morphosyntax reveal basic relations between properties of linguistic objects. Agreement phenomena are instances of co-variation of linguistic forms, which is typically realised as feature congruity, i.e. compatibility of values of identical grammatical categories of syntactically combined linguistic items. Along with government and juxtaposition, co-variation belongs to what (Schmidt and Lehfeldt, 1995) regard as morphological signalling of direct syntactic relations.

Agreement is a relatively well-researched topic, especially in Slavic linguistics (Corbett, 2000 / 1992). Most investigations typically concentrate on the linguistic items themselves (as agreement sources) and on the relevant properties of these items (in terms of agreement features and conditions), while the nature of the relations holding between the 'agreeing' items would not receive any special attention. However, it is the relational nature of agreement that may serve as the basis for a phenomena-driven modularisation of the co-variation affecting the so-called 'phi-features' (prototypically: person, number, gender) in distinct morphosyntactic settings.

In a multidimensional taxonomy, the space of possible agreement relations is derived from a small number of distinctions suitable for immediate formalization. The descriptive device of relational dependency can be utilized to provide a formal framework for encoding these relationships in such a way that the descriptions can be linked to constraint-based grammar formalisms. As argued in (Avgustinova and Uszkoreit, 2003), nothing in the formal apparatus of HPSG excludes the possibility to organise the relations holding in syntactic constructions in a type hierarchy, where type subsumption is interpretable as modelling a continuum from general – and presumably universal – systematic relations to more and still more specific instances of these relations, resulting from admissible cross-classifications.

The figure below illustrates a typology of agreement phenomena as it would result from admissible cross-classifications in a multiple-inheritance hierarchy. The most general type *sys(tematic)-rel(ation)* is associated with two attributes ARG1 and ARG2, and borrowing terminology from HPSG let us assume that the type of their values is *sign*. We can now define a number of relationships among signs. For the sake of illustration we concentrate here on the *covariation* type, which involves distinct linguistic entities ([1] and [2]) with dedicated features ([3] and [4]) that are identified as *covar(iation)-sources* by an associated two-place predicate. The minimal distinction needed for adequate classification of Slavic agreement is the DOMAIN of co-variation, which can be either instant or inferable, and the TRIGER / TARGET configuration, which can be either balanced or asymmetric. The *instant* sub-type is associated with a two-place predicate *local*, while the *inferable* sub-type with a two-place predicate *non-local*. The *balanced* sub-type involves a two-place predicate *x-center* which establishes the situation where neither of the items can be identified as central to the co-variation, in contrast to the *asymmetric* sub-type that is associated with a disjunctive one-place predicate identifying that one of the related items plays a prominent role in the trigger-target configuration. In the latter case a further refinement of the type hierarchy is needed with regard to the dedicated features of the involved items in order to distinguish *loose* from *strict* asymmetry by means of the two-place predicates *unstipulated* and *unidirectional* respectively.

Eventually, we are in a position to define six classes of co-variation phenomena.



Matching as instant balanced co-variation is found between the auxiliaries and the main verb in periphrastic forms. As discussed in (Avgustinova, 1997), the person-number-gender information in Bulgarian analytic (periphrastic) verb forms can be distributed among several components, namely, the main verb itself and a set of auxiliaries functioning as markers to it (3b-c).

Concord as instant loosely asymmetric co-variation is prototypically found within noun phrases, between the adjective and the noun, or possibly between adjectives that modify the same noun (1c).

Cross-referencing as instant strictly asymmetric co-variation holds between the verb and its subject or complement; the same type of co-variation can be assumed between the verbal clitic pronoun cliticized on the verb and the nominal object cross-referenced by this clitic (1a / 2a-b / 3a / 4a).

Correlation as inferable loosely asymmetric co-variation is typically observed in relative clause constructions between the relative pronoun and the noun modified by the relative clause (4b).

Co-reference as inferable strictly asymmetric co-variation holds between an object (or a verbal clitic cross-referencing this object) and the predicative adjective controlled by it; or more generally, between a referential expression, on the one hand, and a co-referent pronoun or a secondary predicative, on the other hand (2c-d).

Accord as inferable balanced co-variation holds between the subject and the complement which are co-dependents of the same verb (1b).

While various structural syntactic configurations in HPSG appear to be relevant for accommodating the different classes of morphosyntactic co-variation, no straightforward correspondences in terms of dedicated phrasal types are readily available.

(1) *Ona* *rastët* *sčastlivym* *rebënkom*.
she.NOM.3SG.F grow.3SG happy.INST.SG.M child.INST.SG.M
'She grows up as a healthy child.' Russian

(a) *cross-referencing* [3SG] ("Ona", "rastët")
(b) *accord* [SG] ("Ona", "sčastlivym rebënkom")
(c) *concord* [SG.M] ("sčastlivym", "rebënkom")

(2) *Decata* *ja* *vidjaxa* *Maria* *maskirana*.
children.3PL ACC.3SG.F saw.3PL Mary.3SG.F disguised.SG.F
'The children saw Mary disguised.' Bulgarian

(a) *cross-referencing* [3PL] ("Decata", "vidjaxa")
(b) *cross-referencing* [3SG.F] ("Maria", "ja")
(c) *co-reference* [SG.F] ("Maria", "maskirana")
(d) *co-reference* [SG.F] ("ja", "maskirana")

(3) *Ti* *si* *štjala* *da* *dojdeš*.
you.2SG AUX2SG AUX.SG.F PRT come.2SG
'You would come (reportedly).' Bulgarian

(a) *cross-referencing* [2SG] ("Ti", "si štjala da dojdeš")
(b) *matching* [2SG.F] ("si" "štjala")
(c) *matching* [2SG.F] ("si štjala" "dojdeš")

(4) *Vliza* *studentyt,* *za* *kogoto* *sega* *govorim*.
comes.3SG student.3SG.M about whom.SG.M now talk.1PL
'The student about whom we are talking now comes in.' Bulgariana)

(a) *cross-referencing* [3SG] ("Vliza", "stydentyt")
(b) *correlation* [SG.M] ("studentyt", "kogoto")

5. Discussion and Outlook

One part of our proposal is to view grammatical structure as a set of systematic relations among syntactically relevant units. Such a perspective exists in linguistic theory as an inherent component of the dependency grammar tradition. Yet, we also hail and exploit the type formalism and other major ingredients of HPSG. The strength of the predominant contemporary constraint-based grammar models HPSG and LFG to a large degree rest on careful combinations of elements from phrase-structure, dependency and categorial grammar traditions.

Adding morphosyntactic relational types in HPSG would result in strengthening the dependency-grammar aspect of the declarative constraint-based formalism with features, multidimensional type hierarchies, inheritance and under-specification.

The other part of our proposal is the systematic approach to the description of shared and non-shared portions of grammar among human languages. We believe that the sketched relational approach is best suited for this purpose since all syntactic phenomena are relational by the basic definition of the concept of syntax. Yet with or without this relational approach, our systematic strategy toward the sharing of grammatical knowledge has applications in several areas of theoretical and applied linguistics.

The most cited purpose is the efficient development of grammars for multiple languages. The shared grammar strategy helps both in parallel multilingual development and in the sequential development of different grammars. Both effects have been described in the DELPH-IN literature.

A further advantage is the use of the shared grammar method and resource in

the development of cross-language and mixed-language applications. The classical cross-language application is machine translation. Others are cross-lingual question answering or cross-lingual IR. Mixed-language applications are able to deal with inputs that mix phrases and sentences from two or more languages as we know it from code switching, foreign-language citations or mixed-language documents.

The most obvious advantages in theoretical research will be in typology. The approach offers a formal foundation for the concrete description of the many-dimensional typology space. Moreover, if appropriately realized, the formalism of the typology would be compatible with the formalism for the description of grammatical knowledge and it would also be suited for computational modelling.

Other potential advantages concern the research on bilingualism and multi-lingualism. The investigation of grammar interference and contamination in bilingual speakers could benefit considerably from a systematic way to isolate and cleanly specify overlap in grammatical knowledge.

On the practical side, such an approach would also facilitate the design of computer-assisted language learning (CALL) software and specialized grammar checkers for non-native speakers.

In addition, it could help in designing tools for machine assisted human translation since the shared grammar approach would help to identify "false friends" above the lexeme level, i.e. in morphology, constituent structure and word order.

All theoretical and empirical investigation of language variation, i.e. in dialectology and sociolinguistics, could utilize formal grammars if these had a natural and systematic method for specifying minimal differences between variants.

The logical next step would then be the exploitation of the approach in the investigation of language change since language change involves the emergence, co-existence and selection of competing minimal language variants.

Summarizing the discussion, we may conclude that the initial results in the description of differences among Slavic languages and the indicated range of possible theoretical and practical benefits of our approach demand a serious continuation of the research, hopefully together with additional companions. The ambitious and demanding goal surely asks for sharing.

References

Avgustinova, Tania. 1997. *Word order and clitics in Bulgarian*. Volume 5: Saarbrücken Dissertations in Computational Linguistics and Language Technology. Saarbrücken: Universität des Saarlandes / DFKI.

Avgustinova, Tania, and Uszkoreit, Hans. 2000. An ontology of systematic relations for a shared grammar of Slavic. In *18th International Conference on Computational Linguistics COLING'2000*, 28-34. Saarbrücken.

Avgustinova, Tania. 2002. *Shared Grammatical Resources for Slavic Languages. Selected topics in multilingual grammar design with special reference to Slavic morphosyntax*, Department of General and Computational Linguistics, Saarland University: Habilitationsschrift.

Avgustinova, Tania. 2003. Metagrammar of systematic relations: a study with special reference to Slavic morphosyntax. In *Syntactic Structures and Morphological Information*, eds. Uwe Junghanns and Luka Szucsich, 1-24. Berlin / New York: Mouton de Gruyter.

Avgustinova, Tania, and Uszkoreit, Hans. 2003. Towards a typology of agreement

phenomena. In *The Role of Agreement in Natural Language: TLS 5 Proceedings*, ed. William E. Griffin. Austin, TX: Texas Linguistics Forum.

Corbett, Greville G. 2000 / 1992. Agreement in the Slavonic Languages: A Provisional Bibliography.

Dörre, Jochen, Eisele, Andreas, and Seiffert, Roland. 1992. Grammars as relational dependencies. Stuttgart: Institut für maschinelle Sprachverarbeitung.

Kameyama, Megumi. 1988. Atomization in Grammar Sharing. Paper presented at *26th Annual Meeting of Association for Computational Linguistics*, Buffalo, New York.

Schmidt, Peter, and Lehfeldt, Werner. 1995. *Kongruenz - Rektion - Adjunktion. Systematische und historische Untersuchungen zur allgemeinen Morphosyntax und zu den Wortfügungen (slovosochetanija) im Russischen*: Specimina Philologiae Slavicae. München: Otto Sagner.