SUBJECT ELLIPSIS IN RUSSIAN AND POLISH*

Marjorie McShane

Abstract. Polish and Russian are partial pro-drop languages with different patterns of subject realization: whereas pronominal subjects are more commonly elided in Polish, they are more commonly overt in Russian. This suggests an approach to modeling ellipsis that begins with the parameter BASELINE SUBJECT REALIZATION with the values OVERT and ELIDED. Once a language has been thus parameterized (OVERT for Russian, ELIDED for Polish), the rules for determining when to override that baseline decision must be formulated. At least four types of factors influence the realization decision for subjects in Russian and Polish: syntactic constraints; the semantic and stylistic nuances of the utterance; and two pragmatic constraints – the necessity of avoiding ambiguity and the preference for avoiding redundancy. While many theoretical approaches would stop at the point of making such generalizations, the approach taken here, which follows the theory of Ontological Semantics, necessitates that all of the proposed constraints be associated with heuristics sufficient to provide a non-native speaker or intelligent agent with a robust, practical knowledge of ellipsis usage.

1. Introduction

This paper presents an analysis of subject ellipsis in Russian and Polish, two languages that can be considered partial pro-drop languages since they show a combination of overt and elided subjects. The languages differ, however, in the prevalence of subject ellipsis, with Polish employing ellipsis in far more contexts than Russian. This distinction suggests an approach to modeling that begins with the parameter BASELINE SUBJECT REALIZATION whose values are OVERT (for Polish) and ELIDED (for Russian). Positing such a parameter raises a number of questions:

1. How can we prove that such a parameter exists? We cannot: as with all other parameters put forth in the literature, we can only test if it is useful in furthering the various models of language being explored by the community.
2. How would such a parameter get set in a child? That falls under the purview of psycholinguistics, which will not be explored here.
3. Why should such a parameter exist? It might be associated with differences in the richness of agreement morphology on the verb and whether or not the language has an overt copula (see below for discussion).

* Thanks to Blazej Bulka for his native speaker judgments and insights; to Anders Holmberg, Sergei Nirenburg and an anonymous reviewer for their constructive critiques of drafts of this paper; and to the developers of the freely available IPI PAN Polish corpus, which proved very useful for this research.

Studia Linguistica 63(1) 2009, pp. 98–132. Published by Blackwell Publishing Ltd.,
9600 Garsington Road, Oxford OX4 2DQ, UK, and 350 Main Street, Malden, MA 02148, USA

4. Under what circumstances is the baseline realization decision overridden? This is the question to which the majority of the exposition is devoted.

Unlike the other contributions to this volume, this paper makes no direct claims about generative syntactic theory. Instead, it advances a computational linguistic theory called Ontological Semantics (Nirenburg & Raskin 2004), which is used for knowledge-rich, reasoning-oriented natural language processing in applications like question answering, machine translation and dialogue between people and intelligent agents. The current analysis along with my previous work on ellipsis (McShane 1998, 2005, etc.) can be thought of as components of the microtheory of ellipsis within the framework of Ontological Semantics. Considering most readers' generative syntactic expectations, this paper begins by describing the motivation for the approach taken here that places it in the context of broader work in linguistics, computational linguistics and artificial intelligence.

1.1. *Ontological Semantics*

Natural language processing based on the theory of Ontological Semantics involves deep semantic (as opposed to light semantic or knowledge-lean) processing of text, with text meaning being expressed using a metalanguage grounded in a language-independent ontology. For natural language (NL) analysis, this means automatically creating a disambiguated text-meaning representation of any input, be it canonically well formed or fragmentary. For NL generation, it means generating natural sounding, context-appropriate utterances – including fragmentary and elliptical ones – from text-meaning representations. Within such a framework, the main research and development involves building computer tractable knowledge bases and reasoning engines that guide the machine in carrying out its many subtasks. For analysis, these include preprocessing (recognizing sentence boundaries, dates, proper names, etc.), full syntactic dependency parsing, word-sense disambiguation, the establishment of semantic dependencies and reference relations, and recovering the meaning not only of syntactically elided categories but also of what we call semantically elided categories. For example, in *Have you begun __ your dissertation yet?*, the action that is elided is most likely *writing*/*working on* although other interpretations are also possible: if everyone were assigned a dissertation to read, the action would be reading, and if everyone were assigned a dissertation to bind the action would be binding (see McShane et al. 2004 for further discussion of semantic ellipsis). A corresponding set of tasks is involved in language generation.

One can see that the linguistic questions driving development in Ontological Semantics are not the ones pursued in generative grammar,

which (in broad terms) seeks to specify the syntactic rules needed to generate well-formed sentences in natural languages. Ontological Semantics is constrained neither to syntax nor to well-formed utterances; in fact, recent interest in the machine processing of blogs and e-mail corpora has brought ill-formed utterances center stage. Instead, the pivotal question in Ontological Semantics is *How can we prepare a machine to extract the same meaning from linguistic utterances as a human would and to generate utterances that will permit a human to readily interpret the machine's intended meaning?* (The machine can, of course, have an intended meaning, in its role as an intelligent agent.) Applied to the field of ellipsis, the Ontological Semantics approach involves not only determining whether a given type of ellipsis is, in principle, permitted in a language as evidenced by a few valid examples, but rather providing the machine with sufficient knowledge and rules to generate and resolve ellipsis with the facility of a native speaker. Developing such resources should involve the analysis of thousands of examples attested in corpora. However, considering the labor-intensive nature such an undertaking, the current study must restrict itself to a subset of phenomena. It is hoped that the theoretical and methodological framing of the relevant questions will encourage others in the field to take up similar analyses with the goal of elucidating the complex rules of ellipsis usage cross-linguistically.

Perhaps the most important aspect of the Ontological Semantics approach is that it does not assume any prerequisites that the system itself cannot fulfill. While one might assume that this should be the case for all computational approaches, it is not. Two examples are (a) Centering Theory, which is a discourse-oriented theory of pronoun resolution for which key concepts like 'utterance', 'realization' and 'ranking' are not sufficiently defined to be automatically detectable (see Poesio et al. 2000 for a critical overview) and (b) the majority of knowledge-lean anaphora resolution systems, which rely on manual preprocessing of the corpus before the reference resolution programs are launched (Mitkov 2001). The Ontological Semantics requirement that all heuristics be machine-tractable with no outstanding prerequisites significantly influences the nature of the analysis presented here, most prominently by not considering accounts based on discourse notions like topic, comment or focus to be the end point of analysis. Analysis at this level was available decades ago in the extensive literature on discourse structure produced in the Slavic linguistic tradition, and while such discourse-oriented analyses are entirely reasonable, they are difficult to implement computationally or to make use of if one does not have the advantage of native-speaker intuitions that suggest the right answer from the outset. So, assuming no native speaker intuitions, how can one detect that a given category is the topic, the comment, or a focused category? By leveraging heuristics from other aspects of the language system.

In theoretical terms, this raises a difficult question. If a syntactic (or phonetic, semantic, etc.) heuristic is sufficient to predict some elliptical phenomenon, is that phenomenon, at base, syntactic (phonetic, semantic) or is it still grounded in discourse considerations that happen to have predictable syntactic (phonetic, semantic) reflexes? The current analysis will not attempt to provide a definitive answer to this question, especially as any answer will inevitably reflect a theoretical bias. Instead, discourse notions will serve as a conceptual substrate for heuristics that can actually be implemented in an environment like OntoSem, which is the practical implementation of the theory of Ontological Semantics.

One current application of OntoSem is a medical simulation and tutoring environment called Maryland Virtual Patient (MVP) (McShane et al. 2007, 2008). The dialogue model under development for this system must support dialogue about medical matters between a physician in training and various virtual agents in the system: the patient, the mentor, medical specialists, etc. This dialogue model relies on plan- and goal-based reasoning in addition to linguistic analysis.[1] I mention MVP because, when its dialogue system has been implemented, the ellipsis resolution module will validate the approach described here, if not the specific heuristics suggested for Slavic.[2] For methods of evaluating the Slavic heuristics, see Section 1.4.

The main static knowledge bases used in OntoSem are an unambiguous, language-independent ontology (world model) and a lexicon for each language processed. The ontology is organized as a tangled tree of inheritance based on the IS-A relation. Concepts are described using property-facet-value triples as well as scripts, which are descriptions of complex events and their participants. Currently scripts cover domains ranging from travel to bankruptcy to disease processes and their treatments. The meaning of words and phrases in a language is recorded in the lexicon, where semantic descriptions refer to elements in the ontology as well as extra-ontological descriptors (McShane et al. 2005a). The meaning of text is recorded using the text-meaning representation language, which is a formal language suited to automatic reasoning.

Although Ontological Semantics can incorporate any kind of machine-tractable heuristics, the environment has not yet been used for speech-based applications, which implies that there is no microtheory of prosody that can be leveraged for the current study. As such, all of the heuristics suggested here relate to aspects of language that can be detected from written texts. Ultimately, however, the model should incorporate prosody as a full-fledged source of heuristics, since even when a person is reading a

[1] Another system in which dialogue and reasoning are closely coupled is TRAINS, which is a conversational planning agent in the realm of cargo transport (see, e.g., Traum 1994).

[2] MVP is currently being developed only for English, although past OntoSem applications have involved Spanish, Chinese, Turkish and Georgian.

text he "hears" its prosodic features, and those features contribute to the interpretation of text meaning. Therefore, it is entirely possible that some of the heuristics suggested here will eventually be superceded by other ones deriving from the prosodic (or some other) module of the language system.

1.2. *Multi-Modal vs. Single-Mode Approaches*

The current analysis incorporates heuristics from many aspects of the language system, including morphology, syntax, lexical and compositional semantics, discourse, stylistics and (when associated text-based clues are available) phonetics. Exploiting hybrid heuristics is not new to science, having been a cornerstone of Artificial Intelligence since its inception in the 1950s; however, only recently has the theoretical linguistics community begun to tackle interface phenomena, which require multi-modal approaches. A robust theoretical infrastructure for treating such phenomena is not yet available, although Optimality Theory might eventually fill this role. Interface phenomena force investigators to abandon the theoretical simplification that linguistic modules like syntax or morphology can be cleanly separated, and raise the question of whether progress in one module will remain valid once that module is incorporated into a comprehensive multi-modal theory or system.

The ellipsis literature of the past 30 years, affected by the dominance of the generative syntactic movement, reveals an overwhelming preference for method-driven (typically single-mode) over data-driven (typically multi-modal) research. In the case of subject ellipsis, this has led to a strong focus on licensing conditions in principle, with less attention paid to conditions on ellipsis usage – which makes sense because the latter falls outside of the purview of generative theory, at least in all cases when the conditions in question are not purely syntactic in nature. However, method-driven approaches suffer from well-known pitfalls, as evidenced by a recent debate in the literature on subsentential utterances, also called fragments. Merchant (2004) argues for an elliptical treatment of fragments and against direct interpretation because direct interpretation would require independent mechanisms for such phenomena as case-marking and the possibility or impossibility of preposition-stranding within a fragment. The drawback of this syntactic bootstrapping approach, however, is that it only works in a subset of cases, as Merchant himself points out and Stainton (2006) underscores in his argument against Merchant's analysis. The real problem is that when syntax fails to predict the correct type or form of a fragment, Merchant falls back on an unspecified semantic recovery process. However, if the mechanisms for semantic recovery are necessary in the language system *anyway*, they must be available to all inputs, not only those structures for

which no syntactic mechanisms are readily available (for further discussion, see McShane et al. 2005b). Purely syntactic studies of ellipsis must be subject to similar scrutiny: will the generalizations put forth hold up when all of the phenomena sidelined as out of purview are reintroduced into a more comprehensive account? This we cannot know until more comprehensive accounts are available. The current analysis takes a step in that direction for subject ellipsis in Russian and Polish.

1.3. *Comparisons with Previous Work*

There are relatively few works comparable to the current analysis. One is Grimshaw and Samek-Lodovici (1998), which treats the overt/elided status of subjects in English and Italian within an Optimality Theoretic framework. One of the central constraints in this analysis is DROPTOPIC: "Leave arguments coreferential with the topic structurally unrealized. Failed by overt constituents which are coreferential with the topic" (ibid: 194). Clearly, this constraint requires specification of the discourse topic – a process that must, however, be carried out manually. Manual annotation of discourse functions is a method-based simplification that can support theoretical work but conceals many of the key aspects of ellipsis usage. The current study, by contrast, while also referring to discourse-oriented generalizations, follows them up with concrete heuristics that indicate when the constraints underlying those generalizations are met.

Within the descriptive Slavic literature most treatments of subject ellipsis are couched in discussions of discourse whose base concepts, like theme and rheme (i.e., topic and comment), are not formalized rigorously enough to make the corresponding entities detectable automatically (see McShane 1998 for references to the Russian and Polish literature on ellipsis).

Within the generative syntactic tradition, the most extensive Slavic contribution is Lindseth (1998), which pursues the question of whether Russian, Czech and Sorbian are so-called "canonical null-subject languages". The discussion centers around three typical characteristics of null-subject languages: in a null-subject language, (1) only null pronominals are stylistically unmarked, (2) only null pronominals can function as bound variables, and (3) only null pronominals can have arbitrary reference (ibid: 34). According to these criteria, Lindseth concludes that most West and South Slavic languages, excluding Sorbian, are canonical null-subject languages (Polish is a West Slavic language that Lindseth does not treat in detail). However Russian (an East Slavic language) is not a canonical null-subject language. The catch to this syntactic treatment, however, is that the first and third characteristics spawn a host of non-syntactic questions. The first characteristic – that

only null pronominals are stylistically unmarked – is theoretically problematic on at least two counts: a) the notion of stylistic marking lies outside of syntactic theory, meaning that ''stylistically unmarked'' is undeterminable; and b) in all canonical null-subject languages, pronouns can be overt without stylistic marking given certain functional considerations, like the need to disambiguate, which also lies outside of the theory. The third characteristic – that only null pronominals can have arbitrary reference – also relies on a semantic notion, arbitrary reference, that is generally considered to be outside the purview of syntactic theory (meaning outside of the purview of what syntactic theory treats in reasonable detail). Thus, the theoretical validity of a *syntactic* class called ''canonical null-subject language'' is less compelling once the largely non-syntactic details underlying the label are scrutinized.

A recent psycholinguistic study by Gordishevsky and Avrutin (2004) pursues the question of whether the widespread optionality in the realization of subject pronouns in Russian slows down acquisition: it does not. They conclude that Russian-speaking children possess, rather early, a subtle knowledge of both syntactic and discourse constraints, and that information structure seems to develop much faster than previously assumed, with parts of it possibly being innate. The fact that information structure might be innate and syntax might be innate opens the question of whether these aspects of the language system can and should be so completely separated in theoretical treatments.

My own previous work on Russian and Polish has concentrated on developing algorithms that can guide both people and machines in making ellipsis decisions (McShane 1998, 1999a,b, 2000a,b, 2005). This body of work conspicuously glosses over subject ellipsis due to my early method-based (generative grammar) orientation. Concentrating on syntax, I sought whatever aspects of ellipsis in Russian and Polish could be accounted for at least in part syntactically, then incorporated heuristics from outside of syntax to explain minimal pairs. Since the employment of subject ellipsis, at least first blush, appears so solidly grounded in discourse considerations, I favored other types of ellipsis. But subsequent work within Ontological Semantics has provided the conceptual framework needed to explore subject ellipsis in a fruitful way.

1.4. *Validation and Evaluation*

Validating the Russian and Polish heuristics presented here could be done in several ways, all of which are labor-intensive. First, one could incorporate these heuristics into existing natural language processing systems for Russian and Polish and see if they improve system performance: for example, if the system is just a parser, is parsing accuracy improved?; and if the system carries out knowledge extraction,

are the precision and recall scores improved?[3] To our knowledge, there do not exist Russian or Polish systems that are semantically rich enough to exploit all of the heuristics presented here, so such an external basis for evaluation would cover only a subset of heuristics.

Second, one could build OntoSem resources for these languages and test the heuristics in the environment for which they were developed. This would involve: (a) building OntoSem lexicons for each language that could be bootstrapped from the current English lexicon; (b) either building from scratch or importing and then extending a parser for each language; and (c) configuring an application (all other knowledge resources and processors being language-independent). Since a small domain would be sufficient for at least an initial evaluation, developing the lexical resources is not a problem, but building the parsers is, since any truly indicative evaluation would need to cover a large inventory of syntactic structures.

Third, one could manually carry out what one envisions the machine should do given an idealized parser, lexicon, ontology, and suite of reasoners. This is, in fact, what was done informally in compiling the heuristics below. The problem with launching a full-scale evaluation of this type can be found in the substantial literature on corpus annotation, which is another domain where manual work stands in for what a machine should ultimately do automatically. Even creating the guidelines for such an effort is always fraught with complexity, but without strict guidelines the validity of the ultimate evaluation would be under question.

One final thing would be needed before embarking on a formal evaluation: an algorithm that incorporates these heuristics into a given language processing system. Without knowing how a particular parser or semantic analyzer is configured, one cannot predict how new rules should be incorporated into it; therefore, formulating a complete algorithm outside of a system is not justified. In short, while the heuristics here assume an NLP environment that has the capabilities of OntoSem, there is nothing system-specific about them and, depending on the capabilities of a given system, the heuristics can be used individually or in any combination.

2. A Typological Comparison of Russian and Polish

Typological features of Russian and Polish that have in some way been linked to ellipsis are shown in Table 1. The features that span the

[3] Precision and recall are typical metrics used for system evaluation: precision measures how many of the returned responses are correct, and recall measures how many of the total number of correct responses the system actually produced.

Table 1. Typological features of Russian and Polish related to ellipsis

Russian	Polish
Rich verbal and nominal inflection	
"Free" word order	
Theme-rheme (topic-comment) discourse structure	
Subject-verb agreement	
No object-verb agreement	
Present and future tense verbal inflection indicates person	
Past tense verbal inflection does not indicate person	Past tense verbal inflection does indicate person
No present-tense form of 'be'	Has present-tense form of 'be'

columns are common to both languages, whereas the ones specific to a column indicate contrasts.

The last two rows contain the contrasts that have been suggested as diachronic explanations for the different patterns of subject ellipsis in the two languages: since Polish verbs show person agreement in all tenses (including the past), elided subjects are more unambiguously recovered, which historically led to the development of a strong pattern of subject ellipsis; by contrast, since Russian verbs not only lack person agreement in the past tense but also have no present-tense form of *be*, over time it developed a system that preferred overt pronominal subjects. (See Lindseth 1998 for more historical discussion.).

Both languages widely permit subjectless sentences that do not represent subject ellipsis, meaning that there is a significantly "subject-less" environment that could influence the prevalence of subject ellipsis. Types of subjectless sentences that are not elliptical include unagentive impersonal verbs, derived impersonal verbs, non-verbal impersonal words (like Russian *nado* and Polish *trzeba* 'it is necessary/one must'), non-finite verb forms used as a main verb, the 2nd person generalized human construction, the indefinite personal (3rd person plural) construction, agentive impersonal constructions (Polish only), and the omission of all relevant arguments in dialogue strategies (e.g., a yes-no question can be answered by repeating the verb alone, with or without negation). Descriptions and examples of all of these phenomena can be found in McShane 2005:204-230. The prevalence of these strategies is especially important for interpreting the increasingly common corpus studies of language phenomena, since when things are counted, it is essential to understand *what* is being counted. The case in point: counting sentences that lack a nominative-case subject would give a very skewed view of the elliptical picture of Russian or Polish unless the corpus were

carefully pruned to include only those sentences that would otherwise permit an overt subject.[4,5]

Both Polish and Russian show widespread ellipsis of other categories, including objects, head nouns, conjunctions, particles and, for Polish only, inflectional morphology (McShane 1998, 2005). (Naturally, these languages permit verbal ellipsis as well, but so do many non pro-drop languages.) These other types of ellipsis are important to keep in mind for two reasons. First, the ellipsis decision for one category very often depends upon the ellipsis decision for another category, a phenomenon that can be referred to as dependencies in ellipsis (McShane 2000a). Second, it is possible that subject ellipsis is not a separate phenomenon but, rather, part of a larger elliptical system, as has been suggested for Russian by various investigators (Franks 1995; Lindseth 1998, who essentially follows Franks' analysis; and numerous others before them). This analysis of Russian seems right, but what, then, should one make of Polish? It certainly has more frequent subject ellipsis and less frequent object ellipsis than Russian, but is frequency sufficient to defend a theoretical split? And if so, then where is the frequency cutoff?

The prevalence of subject ellipsis in Russian and Polish is affected, among other things, by whether the genre is written language or spoken language, as shown in Table 2.

Two reasons are typically cited for why, in Russian, the spoken language permits more subject ellipsis than the written language: (1) the presence of discourse cues, intonation, and other aspects of colloquial style support the recovery of elided material; (2) ellipsis can add a desired

Table 2. Gross generalizations about subject realization in Russian and Polish

	Written	Spoken
Russian	The subject tends to be overt	Considerably more ellipsis than in the written language
Polish	The subject tends to be elided	Considerably more overt subjects than in the written language

[4] A number of contributions offer counts of overt and elided subjects in Russian and Polish. Seo (2001) compared subject usage in five Slavic languages but found his method of comparing translations of literature to be fraught with confounding factors, particularly morphosyntactic variation in translations that makes direct multi-lingual comparison difficult. Nilsson (1982) also provides some counts, essentially showing that even strongly elliptical Polish shows many overt subjects.

[5] Studying subject ellipsis in Polish has the significant benefit of permitting the interesting cases – those with an overt pronoun – to be searched for in corpora. By contrast, for languages like Russian, where subject ellipsis is less prominent, the elliptical cases need to be collected by hand since current parsers are not robust enough to detect the gaps with a high degree of confidence.

Table 3. Overview of subject ellipsis in Polish and Russian

Language	Baseline realization option	What can overrule the baseline realization
Polish	elided	1. syntactic configuration (syntax) 2. potential for ambiguity (semantics/pragmatics) 3. emphasis (semantics)
Russian	overt	1. syntactic configuration (syntax) 2. avoidance of redundancy (pragmatics) 3. avoidance of long-winded descriptions (semantics/pragmatics) 4. stylistic nuances (stylistics)

colloquial flavor to the speech; but note that it does not always have this stylistic effect – it is often a stylistically neutral choice. The typical justification for why the Polish spoken language includes more overt subjects is to permit the speaker to convey emphasis and emotional flavor.

It is common in the linguistic community to make hard divisions between modules of the language system. Although clean lines cannot always be drawn, this work reflects an attempt to make such distinctions in order to make the analysis most relevant to the largest number of readers. Table 3 provides a high-level overview of subject ellipsis in Polish and Russian, with the relevant language modules indicated.

It must be emphasized that this categorization of phenomena is one of many possible categorizations, having been selected to serve practical goals. As mentioned above, it is often unclear how to separate the reason for something from its surface manifestation: e.g., are syntactic aspects of ellipsis really syntactic in nature, or do they always derive from pragmatic considerations? The latter would seem to be theoretically more satisfying but in practical terms would add a layer of analysis that cannot be proven to be necessary. For example, both Russian and Polish do not permit an overt 2nd clause subject in configurations like *He$_i$ left work and *he$_i$ drove home*. This can be framed as a simple syntactic rule but it arguably derives from a pragmatic constraint regarding the felicity of repeating a, by necessity, thematic preverbal subject in the latter conjunct of a coordinate configuration. Whether or not we need to refer to the underlying pragmatic constraint depends upon the theoretical framework or practical application in question.

Table 3 does not include the heuristics for how to meet the conditions for overriding the baseline realization: those are discussed in the subsections below. It is noteworthy that the relevant heuristics do not

necessarily belong to the same module of the language system as the high-level constraint in question: for example, the potential for ambiguity is a semantic or pragmatic consideration – one wants the interlocutor to correctly understand one's meaning – but the heuristics that suggest when such ambiguity might arise are phonetic, semantic and syntactic.

3. Predicting Subject Ellipsis in Polish

In both written and spoken Polish, subject pronouns tend to be elided, but in the spoken language, overt realization is also often possible. As Nilsson (1982:58) writes: "In colloquially colored Polish texts, dialogues, etc. explicit pronominal subjects sometimes have no special communicative function. They are not even expressive. Deletion of the subject would neither cause any confusion as to its identity nor alter the communicative structure of the sentence. In other words, the pattern is similar to the Russian one". Cases of true optionality are quite difficult to account for and are not yet satisfactorily covered in the ellipsis model presented here.

The subsections below are all rejoinders to **In Polish, pronominal subjects are elided unless**…, with the natural implication being that all the sentences in question would have a pronominal subject, be it overt or elided. The determination of whether the referring expression should be pronominal in the first place is another central issue, but one lying outside the scope of this paper.

The classification considers subject ellipsis from the point of view of a non-native speaker or an intelligent agent attempting to fully interpret or optimally generate Polish text. Assuming that the baseline rule is "elide pronominal subjects", different questions arise for language understanding and language generation.

- **Language understanding**: Does the overt status of the subject (if it, in fact, is overt) imply emphasis, or is there a non-semantic explanation for why the subject is overt? In other words, should the text-meaning representation that an intelligent agent creates for the input include emphasis or not?
- **Language generation**: Given the meaning to be expressed – which, for an intelligent agent within OntoSem will be recorded in a text-meaning representation – should an overt subject be used to (a) avoid ambiguity, (b) convey the emphasis that is reflected in the text-meaning representation, (c) fulfill some syntactic constraint?

Examples throughout are long and therefore are translated but, in most cases, not glossed. In some cases, grammatical information is provided on key words. The overt subject of interest is in boldface and is always in the nominative case. Elided subjects are indicated by [e], with the elided pronoun being indicated in the translation: [he]. Sources of

citations are indicated using abbreviations whose full references are provided at the end of the paper.

3.1. *Syntactic Configurations Requiring an Overt Subject*

The configurations in this subsection reflect hard syntactic rules aimed at avoiding word salad.

3.1.1. *The Pronoun is Modified*

Modification of a pronoun requires it to be overt. Among the modifiers in Polish that can modify pronouns are *i* 'even', *nie* 'it isn't/wasn't (someone)', *to* 'it was/is (someone)' and *też* 'too, as well'.

(1) Milczenie trwało tak długo, że wreszcie i **ja** spojrzałam. (IPI)
'The silence continued for so long that finally even **I** looked.'

(2) Nie **ja** tę wojnę zacząłem. (IPI)
not I_{NOM} $this_{ACC}$ war_{ACC} started.
It wasn't me who started this war.

In the following example, note that in Polish, verbal morphology can cliticize onto a pronominal subject. Here, the 1st person plural ending *śmy* moves from the verb *wysłali* 'sent$_{3.PL}$' to the subject pronoun *my* 'we$_{1.PL}$'

(3) Kto nam dowiedzie, że to **my**śmy ich wysłali? (IPI)
'Who will tell us that it is **we** (who) sent them?'

Another way of modifying a necessarily overt pronoun is to provide it with a parenthetical description like the following:

(4) Wielu z nich mogłoby zginąć, a **ja**, ich komendant, jestem za nich odpowiedzialny. (IPI)
'Many of them could have perished but **I**, their commander, am responsible for them.'

Pragmatically speaking, in cases of pronominal modification the pronoun must be overt because it is in focus. Its overt status does not convey any additional semantics or pragmatics – the subject simply hosts its modifier, with the modifier carrying the additional meaning.

3.1.2. *The Verb is Elided*

Polish employs all of the cross-linguistically prevalent types of verbal ellipsis (Lobeck 1995) as well as some less common types (McShane 2000b). Some of these, like gapping (5) and main verb ellipsis licensed by two arguments or adjuncts (6), require an overt subject in order to create a clause, with the overt status of the subject carrying no special implications.

(5) Omar należy do plemienia Hottak, **ja** [e] do Stanakzajów. (IPI)
'Omar belongs to the Hottak tribe, **I**, [e] to the Stanakzajs.'

(6) „Ona stawia czajnik: «Zimno». **Ja** [e] nic" (Pok: 341).
'She puts on the kettle: "It's cold." **I** don't respond/don't do anything.'
[lit: $\boldsymbol{I}_{NOM}$ *[e] nothing*$_{ACC}$*]*

3.2. *An Overt Subject is Needed Due to Potential Ambiguity*

One baseline constraint on ellipsis is, Do not elide if ellipsis can result in ambiguity. This section presents a series of configurations that, without an overt subject, could result in ambiguity and/or cause difficulty in establishing the correct reference link. Using an overt subject in these configurations has no special semantic implications.

Below is a summary of the relevant heuristics, the language modules to which they belong, and the subsections in which they are discussed. To reiterate a point made earlier, the fact that the heuristics derive from various aspects of the language system does not change the fact that the potential for ambiguity – a semantic/pragmatic factor – is driving the use of the overt subject in all cases.

3.2.1 The antecedent is insufficiently syntactically accessible to guarantee recoverability of the elided subject [*syntax*]
3.2.2 The third person subject of a subordinate clause is not coreferential with the subject of the matrix clause *[morpho-syntax]*
3.2.3 There is a shift in the subject (a syntactic heuristic) or the agent/experiencer (a semantic heuristic) in sequential main clauses *[morpho-syntax and/or semantics]*
3.2.4 There is a shift in agent/experiencer between a denominal and a subsequent tensed clause *[morpho-syntax & semantics]*
3.2.5 The antecedent is a rhematic subject as detected by word order *[morpho-syntax]*
3.2.6 The personal ending of the verb form is phonetically ambiguous *[morphology & phonetics]*

3.2.1. *The Antecedent is Insufficiently Syntactically Accessible to Guarantee Recoverability of the Elided Subject*

Accessibility hierarchies are a fruitful conceptual tool, with syntactic hierarchies typically ranking the subject first, followed by the direct object, indirect object, and various oblique objects and objects of prepositions. The relevance of antecedent accessibility to ellipsis was shown in McShane (1998, 2005) with respect to object ellipsis potential in Russian and Polish.

According to the Polish data compiled so far, antecedents occupying at least three local syntactic positions can be unable to support subsequent subject ellipsis: objects of prepositions (7), nominal complements (8), and nominative-case components of copular constructions (9) – (10). In such configurations, the overt status of the subject has no semantic implications.

(7) Szkoda, że [e] nie wiedziałem. [e] Przyszedłbym. Bez Nery_{GEN}, oczywiście, bo **ona** nie rozumie po polsku. (IPI)
'It's too bad that [I] didn't know. [I] would have come. Without Nera_{GEN}, of course, since **she** doesn't speak Polish.'

(8) $\text{Zadanie}_{\text{NOM}}$ $\text{odbiorcy}_{\text{GEN}}$ jest znacznie trudniejsze: musi **on** rozpoznać konkretną wartość każdego słowa.
(Jodłowski 1977:156; quoted from Nilsson 1982:41–42).
'The task_{NOM} of a $\text{customer}_{\text{GEN}}$ is much more difficult: **he** has to evaluate the concrete worth of every word.'

(9) $\text{Czym}_{\text{INSTR}}$ jest $\text{akceptacja}_{\text{NOM}}$ i co **ona** oznacza? (IPI)
'$\text{What}_{\text{INSTR}}$ is $\text{acceptance}_{\text{NOM}}$ and what does **it** mean?'

(10) [e] Dawniej myślałem, że moja_{NOM} żona_{NOM} $\text{Gracja}_{\text{NOM}}$ jest tą_{INSTR} $\text{femmefatale}_{\text{INSTR}}$. Ale **ona** wyjechała, a mnie dalej idzie jak po grudzie. (IPI)
'[I] long thought that my_{NOM} wife_{NOM} $\text{Grace}_{\text{NOM}}$ was a $\text{femme fatale}_{\text{INSTR}}$. But **she** left and it's been tough on me.'

It is noteworthy that in none of the cited examples would a human reader detect any ambiguity, in the sense of competing potential antecedents. This suggests that this constraint represents a syntactically oriented, systematic means of avoiding ambiguity that does not require analysis of the potential for ambiguity in each specific context.

3.2.2. *The Third Person Subject of a Subordinate Clause is Not Coreferential with the Subject of the Matrix Clause*

In Polish, when the subject of a subordinate clause is coreferential with that of the matrix clause, the subordinate clause subject is elided. If the subject of the subordinate clause is *different* from that of the matrix clause, and if both clauses have a 3rd person subject, and if the verb forms do not provide a morphological means of showing that their subjects cannot be coreferential (e.g., by being in the past tense and differing in gender), then the subject of the subordinate clause must be overt – otherwise referential ambiguity would ensue. The main clause subject is, itself, often elided.

(11) a. [e] Wie, że żyje.
[e] *know* 3-SG.PRES *that live* 3-SG.PRES
'[He/she$_i$] knows that [he/she$_i$] is alive'
b. [e] Wie, że on żyje.
[e] *know* 3-SG.PRES *that he* NOM *live* 3-SG.PRES
'[He/she$_i$] knows that he$_k$ is alive.'

3.2.3. *There is a Shift in the Subject Referent of Sequential Main Clauses*

Sequential main clauses can be connected in various ways. Viewed syntactically, they might be linked by an overt coordinating conjunction, an elided coordinating conjunction, a comma, semi-colon, colon, dash or sentence break. Viewed semantically, they can represent a sequence of actions, the second can elaborate on the first, etc. Irrespective of the syntactic or semantic relationship between clauses, the following generalization appears to hold for Polish: when two sequential tensed clauses have different subjects, the second one – which is necessarily rhematic (cf. 3.2.4) – typically can or must be overt. The subject of the first clause may or may not be overt. Unlike the heuristic in 3.2.2, this generalization is not limited to 3rd person subjects. For example, in (12) the subject changes from the implied *you* of the preceding imperative clauses to the overt *I* of the final clause.

(12) – Pomóż mi! … Popatrz gdzie popadnie. Na kredensie zobacz i pod spodem, **ja** zobaczę w pokoju. (Chm1)
'Help$_{IMPER}$ me!... Look$_{IMPER}$ anywhere. Look$_{IMPER}$ on the sideboard and underneath, **I**'ll look in the other room.'

(13) Nathan zaproponował mnie, Arturowi i Robertowi małą wycieczkę do Kanady. Artur nie mógł, a **my**śmy skorzystali. (IPI)
'Nathan offered Arthur, Robert and me a little trip to Canada. Arthur couldn't go but **we** accepted the offer.'

(14) [e] Pamiętam, że [e] czekaliśmy w kolejce na koncert Rubinsteina i **ona** przeszła koło mnie i powiedziała: Musi być trudno z panem flirtować… (IPI)
'[I] remember that [we] were waiting in line at a Rubinstein concert and **she** passed by me and said: It must be hard to flirt with you…'

3.2.4. *There is a Shift in Agent/Experiencer Between a Denominal and a Subsequent Tensed Clause*

Nominalization is very widely used in Polish, and while nominals cannot have a subject (a syntactic entity), they can have an agentive or

experiencer case-role (a semantic entity), which tends to correspond to the subject in syntax-semantics mappings. If the agent or experiencer of a denominal – be it overt or implied – is different from the agent or experiencer represented by the subsequent subject, that subject can be overtly realized. In (15) the agent of the denominal is implied – "observing (by oneself) one's own operation…" and the next subject, which has a different referent, is overt.

(15) Przyglądanie się własnej operacji wymaga ogromnego samozaparcia, a **ja** w owej chwili byłam przygnębiona i rozgoryczona Twoim postępowaniem i absolutnie nie mogłam się zdobyć na dostateczną siłę ducha. (Chm2)
'Observing one's own operation requires a great act of will and **I** at that time was depressed and disillusioned by your behavior and absolutely couldn't muster up sufficient strength.'

It remains to be determined which factors determine whether the latter subject simply can be overt or must be overt.

3.2.5. *The Antecedent is a Rhematic Subject as Detected by Word Order*

When a subject, S1, has an antecedent, S2, that, itself, is a rhematic subject, S1 may optionally be overt with no special semantic implications.

(16) Wobec tego przyjechali rewidenci z powiatu, a potem z województwa. Ustalili (**oni**), że Maria P. sama nie mogła popełnić nadużyć. (Urban, 31).[6]
Besides that came$_{3.PL}$ *auditors*$_{NOM}$ *from district*$_{GEN}$ *and the from province*$_{GEN}$ *Established*$_{3.PL}$ *(they*$_{NOM}$*) that…*
'Besides that, auditors came from the district and then from the province. (**They**) established that Maria P herself could not have committed fraud.'

From the perspective of discourse structure, the reason why an overt subject can be used in such configurations derives from the basics of theme-rheme structure: the theme is what is most expected to be continued in the next utterance; and, if the theme is continued, it is readily elided. By contrast, a continuation of the rheme is less expected, meaning that expressing it overtly can help to establish the necessary reference relation.

Although the formulation of this heuristic in terms of theme/rheme structure is the most compact for descriptive purposes, it does not provide heuristics for detecting which constituent is rhematic. In a

[6] Quoted from Nilsson (1982:41). In the original, the subject is overt; Nilsson makes no mention of whether or not it could be elided. My informant said that it could be elided and added that the word order 'ustalili oni' used to be not accepted as proper Polish but now people are using it more often.

text-only approach (i.e., no intonation), word order is the central clue: typically, thematic elements in Polish come first in the clause, followed by rhematic ones, assuming the neutral intonation typical of written texts. In (16) the subject headed by *auditors* appears after its verb, indicating its rhematic status. Since it is rhematic, the next reference to it can be realized by an overt pronoun.

3.2.6. *The Personal Ending of the Verb Form is Phonetically Ambiguous*

In some cases, the potential for phonetic ambiguity can lead to the use of a disambiguating subject pronoun. For example, 3rd person present tense verb forms ending in *–e* and 1st person present tense verb forms ending in *–ę* sound identical in fast speech. Therefore, in texts conveying spoken language, like (17), the subject can be overt simply to disambiguate, with no emphasis implied.

(17) - Bo fakt, że [e] z nim sypiasz, to żadna zgryzota, [e] rozwiedziony jest.[7]
- **On** chce wyjechać. Do Tybetu. (Chm4)
'The fact that you're sleeping with him isn't a problem, [he] is divorced.' '**He** wants to go away. To Tibet.'

To leverage this generalization in a formal model, one needs to create an inventory of verb forms that potentially sound similar, which should reduce to a few rules covering a few inflectional paradigms. Then the associated rule for subject realization could be formulated as follows: If the verb form is among those with phonetically ambiguous forms and if the given utterance occurs in direct speech and if in the list of candidate antecedents there are ones representing both of the given readings[8] then the subject pronoun should be overt.

3.3. *The Sentence is Emphatic (but Not Contrastive)*

Consider the sentence *I don't want to go* as pronounced in two different contexts:

(a) A couple is sitting by the seaside and the husband asks the wife if she'd like to walk down to the water with him to cool off her feet.
(b) The same couple, 3 minutes later: the husband is still trying to convince the wife that she should go dunk her feet.

[7] In this clause, if the subject were overt the word order would be *on jest rozwiedziony*.
[8] Computational approaches to reference resolution typically create a list of potential antecedents, within a certain window of previous context, then employ various heuristics to select the most probable one.

In situation (a), the wife likely responds with neutral intonation, whereas in (b) her intonation will certainly be quite different – perhaps *want* is stressed, or each word is pronounced unusually distinctly and emphatically. In other words, in (b) the entire sentence *I don't want to go* reflects emphasis – not contrastive emphasis of a given element, but overall emphasis.

This example does not provide a formal definition of emphasis – a task that would have required too much detail for this article – but it is sufficient to convey to an English-speaking audience an ellipsis-related phenomenon in Polish. If an entire utterance in Polish in pronounced emphatically (meaning that emphasis scopes over the entire utterance, whatever that might mean prosodically), this implies one of many possible non-neutral speaker attitudes. The subject of such an utterance can typically be overt, with no implication that the subject itself is emphasized, contrastive, etc. The most easily accessible indicators of this kind of emphasis in written texts are exclamation points, and sentences with exclamation points often have an overt subject that is not licensed by any of the means described above.

(18) Po kilku zakrętach i przejściach faraon znowu odezwał się: - Ależ **my**śmy tu już byli, bodaj że ze dwa razy!... (IPI)
'After a couple of turns and alleyways the Pharaoh again says: ''But **we**'ve already been here, twice!...'' '

(19) Etykieta jednak zobowiązywała do tego, żeby na każde wejście i wyjście królowej grano fanfary. - Ile **my**śmy się za nią nabiegali! (IPI)
'Etiquette required that a fanfare was played every time the queen went in or out. ''**We** had to run after her so much!'' '

Many sentences are semantically such that they can be used either emphatically or in a stylistically neutral way. In the set of examples below, if the subject is overt it implies emphasis scoping over the sentence – which might reflect dissatisfaction, frustration, etc. The glosses show the contrast using italics to indicate words that would be stressed in English. (Recall that boldface is used only to show the subject in question, it does not indicate emphasis.)

(20) O czym **(ty)** mówisz?
'With elided subject: What are **you** talking about?'
'With overt subject: What are **you** *talking* about?'

(21) To pytanie, które mi często stawiają, między innymi zadała mi to pytanie pani minister Jakubowska: dlaczego **(ty)**ś z tym nie przyszedł do mnie? (IPI)
That question that I'm often asked was among the questions that Minister Jakubowska asked me:
'With elided subject: Why didn't **you** come to me with this?'
'With overt subject: Why didn't **you** *come* to me with this?'

(22) - Gdzie te cholerne klucze?... Ty, popatrz, tu gdzieś powinny być klucze, cały pęk. **(Ja)** się śpieszę. Bez kluczy przecież nie wyjdę. (Chm1)
"Where are those blasted keys?... You, look, there must be keys here, a whole ring of them.
With elided subject: '**I**'m in a hurry.'
With overt subject: '**I**'m in a *hurry*.'
But without the keys [I] won't leave.

(23) Na pomost wbiegł Alojz ze spinningiem w ręku. Lekko chwiejąc się na nogach, podszedł do uwiązanej łodzi. – Gdzie (**ty**) się wybierasz? (IPI)
Alojz ran to the foot-bridge with a fishing rod in his hand. Staggering a little, [he] walked up to the tethered boat.
With elided subject: 'Where are **you** going?'
With overt subject: 'Where are **you** *going*?'

It would be convenient for our model if it were the case that all overt subjects in Polish that are not accounted for by the heuristics listed above implied that the entire utterance was emphatic. This, however, is not the case, as many contexts – especially in the spoken language – permit an overt or elided subject with no detectable semantic differences. (24) is one such context.

(24) – Bardzo dobrze. To teraz od razu [e] zadzwonisz do milicji. Do tego kapitana, który się tym zajmował, **on** się nazywa Różewicz. (Chm3)
'Very good. So now [you] will call the police, that captain that was handling this, his name is [lit: **he** self calls] Różewicz.'

(If the subject were elided in (22) – (24), the word order would be different: the verb would precede the particle *się*.)

To recapitulate this section: If a sentence ends with an exclamation point, the utterance is undoubtedly emphatic and an overt pronominal subject can be used. If, however, the sentence does not end with an exclamation point, the presence of an overt subject might, itself, be implying emphasis or it might not. Such optionality will create problems for our intelligent agent in the task of language understanding, since it will need to decide whether to attribute the semantics of emphasis to sentences that contain an overt subject but no other detectable indicators of emphasis. This judgment will require a sophisticated level of semantic analysis and might also be informed by prosodic clues if the agent is processing speech and not text. Subject optionality will not, however, create problems for most practical language generation systems, which need only to generate one acceptable string in each context (in the case of Polish, the default rule of eliding the subject would work in a case like (24)). However, if the ultimate goal is to endow intelligent agents with

language abilities rivaling those of a person, they should be aware of all of their generation options in all contexts.

4. Predicting Subject Ellipsis in Russian

In Russian, the baseline choice for subject realization is ''realize overtly'', and the main goal of the current model – and challenge for an intelligent agent – is to determine when to elide, since not eliding in the appropriate contexts can lead to ungrammaticality or stylistic infractions. The types of factors determining when one can or should elide the subject in Russian are 1) the syntactic configuration; 2) the desire to avoid redundancy; 3) the desire to avoid long-winded formulations, and 4) the desire to convey stylistic nuances.

Many of the configurations that promote subject ellipsis in Russian also promote object ellipsis, as shown in McShane 1998 and 2005. In fact, knowing what to look for as heuristics for subject ellipsis derived largely from work on object ellipsis – again raising the question, Are we dealing with different processes for the ellipsis of different categories or just one generalized process aimed at avoiding the unnecessary repetition of sentence constituents?

The subsections below are all rejoinders to **In Russian, pronominal subjects are overt unless**…, with the natural implication being that all the sentences in question would have a pronominal subject, be it overt or elided. In the examples, the elided subject of interest is represented by [e] and the English translation shows the associated pronoun in brackets [He].

4.1. *Syntactic Configuration*

In Russian coordinate configurations in which the clauses are not complicated by long subordinate clauses, adverbials, etc., the subject of the second clause should be elided: not eliding it would be very strange, if not ungrammatical.

(25) On_i sel i [e]/*on_i načal kušat'.
'He sat down and [e]/*he started to eat.'

However, if the conjuncts are significantly complex, as in (26), then the latter subject can be overt, presumably to make the structure easier to interpret.[9]

[9] This phenomenon is reminiscent of heavy NP shift in English, which permits certain long NPs to move to the right of their typical syntactic positions:

a) I sent to my mother the series of letters I had received from the student loan company.
b) I sent the letter to my mother. / *I sent to my mother the letter.

(26) On sel rjadom so mnoj na divane nedaleko ot kamina i posle dolgogo molčanija, vo vremja kotorogo ja stala čuvstvovat' sebja vse bolee neujutno, (on) načal govorit'.
'He sat down next to me on the couch near the fire and, after a long silence during which I started feeling more and more uncomfortable, (he) started talking.'

Here there is still some preference for the second subject to be elided, but the overt-subject variant is far more acceptable than in (25). If we made the sentence even longer, with more intervening subordinate clauses and modifiers, the felicity of the overt subject would increase still further.

The possibility of having an overt second subject in at least some such configurations raises two connected questions:

- When the second subject is not overt and cannot be overt, as in (25), is the configuration VP-conjunction or clausal conjunction with ellipsis of the second subject?
- When the second subject is not overt but could be overt, as in (26), is the elliptical variant VP-conjunction and the overt-subject variant clausal conjunction (i.e., a native speaker would be evaluating two different parse trees), or are both variants clausal conjunction with a different pragmatic choice being made that accounts for the different subject realizations?

The issue is whether there is a syntax → pragmatics pipeline in processing (in which case the clausal conjunction strategy would always have to be used, since the determination of whether the clauses were sufficiently complex could only come later) or whether pragmatic decisions influence syntactic structures (in which case VP-conjunction could account for the conjunction of simple clauses whereas clausal conjunction would be used for complex ones).

Compare same-subject coordinate structures with same-subject subordinate structures: in the latter, the second subject can be overt without requiring so much "padding" of each clause. For example, whereas (25) is bad with an overt second subject, (27) is significantly less bad, and making the clauses more complex would make it quite acceptable (cf. Section 4.2.1 below).

(27) On sel potomu, čto (on) ustal.
'He sat down because (he) was tired.'

4.2. *Avoidance of Redundancy*

In Russian, having topic status is not a sufficient condition for the subject to be elided. As such, a constraint like Grimshaw and Lodovici's DropTopic constraint is too strong for Russian, at least in the absence of

other constraints that favor overtly expressed subjects. However, there are configurations that underscore the topic status of the subject, making it, in a sense, so markedly topical that eliding it to avoid redundancy is the favored option. The avoidance of redundancy is a pragmatic constraint, but the heuristics that flag redundancy belong to various aspects of the language system:

4.2.1 Same-subject subordinate structures *[syntax]*
4.2.2 A series of three or more actions with the same subject *[syntax]*
4.2.3 Repetition structures *[lexical syntax]*
4.2.4 Independent nominal topic + sentence *[syntax]*
4.2.5 The entity's non-subject antecedent is, itself, elided *[syntax]*
4.2.6 Elaboration *[syntax & semantics]*
4.2.7 "Convenient" ellipsis configurations *[semantics]*

4.2.1. *Same-Subject Subordinate Structures*

When the subject of a subordinate clause has the same referent as that of the matrix clause, the subordinate clause subject is typically elided, although it does not need to be. For example, in (28) - (29) the elliptical version is slightly more natural but the overt-pronominal version would be acceptable as well.

(28) Doroga ix ne ustraivala, poskol'ku [e] ne vela k morju. (Chm5)
'The road didn't' suit them because [it] didn't lead to the sea.'

(29) Ona vybegala i ne lajala, poskol'ku [e] byla sderžannoj sobakoj. (Tok: 156)
'She would run out but not bark since [she] was a well-behaved dog.'

It is worth mentioning in passing the structure of subjunctive clauses in Russian, which are a type of subordinate clause. As discussed in Avrutin and Babyonyshev (1997), Russian shows subjective obviation: the requirement that a pronominal subject of a subjunctive clause be non-coreferential with the matrix subject (30). Subjects of subjunctive clauses cannot be elided in Russian.

(30) Lina_{i} xočet, čtoby $\text{ona}_{j/*i}$ vyigrala.
$\text{Lina}_{\text{i-NOM}}$ $\text{wants}_{\text{3.SG.PRES}}$ that $\text{she}_{\text{j/*i-NOM}}$ $\text{won}_{\text{3.SG.FEM.PAST}}$
'Lina wants her to win.'

If the understood subject of the subjunctive clause is coreferential with the matrix clause, then a necessarily subjectless infinitival construction is used:

(31) Lina xočet vyigrat'.
Lina_{NOM} $\text{wants}_{\text{3.SG}}$ $\text{win}_{\text{INFIN}}$
'Lina wants to win.'

4.2.2. *A Series of Three or More Actions with the Same Subject*

Russian subjects can be, and often should be, elided in the latter clauses/sentences of a series of actions. A discourse-oriented explanation is that the subject becomes thematic in the first action and remains so with the introduction of new rhematic actions. The heuristics that can operationalize this generalization can be framed syntactically: If the subjects of the three or more sequential clauses – which can be separate sentences – have the same referent, then the subjects of the latter ones can, and often should, be omitted, as in (32)–(33).

(32) Togda ja sel v tramvaj i dolga exal, deržas' za ramu ot slabosti i dyša na zamerzšee steklo. [e] Priexal tuda, gde žil Rudol'fi. [e] Pozvonil. Ne otkryvajut. [e] Ešče raz pozvonil. (TR)
'Then I sat in the tram and rode a long time, holding on to the rail from weakness and breathing on the frozen window. [I] arrived at Rudolfi's place. [I] rang. Nobody answered.
[*lit:* NEG open$_{3.PL.PRES}$] [I] rang again.'

(33) Ja, neizvestno začem, položil rjadom s soboju knižku žurnala; s cel'ju čitat', nado polagat'. No [e] ničego ne pročel. [e] Xotel postavit' ešče raz termometr, no [e] ne postavil. (TR)
'I, for some reason, put down beside me an issue of a journal – to read it, apparently. But [I] didn't read anything. [I] wanted to take my temperature again but [I] didn't.'

Orienting this heuristic around three or more sequential clauses optimizes the balance of predictive power and coverage, at least for the data incorporated into this early stage of the model. That is, if one reduces the number of clauses/sentences to two, then in many cases the latter subject is preferably not elided, so predictive power suffers; and if one increases the number of clauses/sentences to more than three, then the number of contexts covered by the heuristic is greatly reduced.

It is noteworthy that in (32) the interference of an indefinite personal sentence (i.e., a sentence with a 3rd person plural verb that has indefinite reference) does not reduce the topicality of the preceding subject, as is shown by the fact that the subject of the final clause can be elided. Two additional ellipsis-promoting factors likely contribute to this ellipsis decision: the final clause is a repetition structure, meaning that a previous verb is repeated (cf. Section 4.2.3), and the subject in question is *I*, which has strong discourse prominence (cf. Section 4.2.8). This is a particularly good example of how a confluence of ellipsis-promoting factors can be involved in an ellipsis decision.

4.2.3. *Repetition Structures*

Repetition structures are what I call contexts in which two consecutive conjuncts or clauses contain the same verb selecting the same object(s) (McShane 2005). In all Russian repetition structures—barring those instances in which overt repetition of arguments serves a stylistic purpose—all arguments should be elided, since the purpose of such structures is to elaborate on the verb, often by adding modifiers as in (34)–(35).

(34) Tretij den' ja ždala zvonka. [e] Zrja ždala. (Chm6: 15)
'For the third day I waited for the call. [I] waited in vain.'

(35) Pered glazami vse vremja, kak navjazčivyj refren, prokručivaetsja mgnovenie, kogda Sereža letit i padaet. [e] Letit dovol'no nelepo. I [e] padaet očen' tjaželo. (Tok: 149).
'Before my eyes all the time, like a never-ending refrain, repeats the moment when Seryozha flies and falls. [He] flies pretty awkwardly. And [e] falls very hard.'

Repetition structures can be detected based on lexico-syntax (as well as, of course, reference resolution): if the same verb with the same overt or implied arguments repeats in two successive clauses, the structure is a repetition structure. Note that it is not only subjects and objects that can be elided in repetition structures but minor parts of speech as well (McShane 1998, 2005).

4.2.4. *Independent Nominal Topic + Sentence*

Presenting an entity as a nominal topic outside of regular clausal structure gives it strong topic status and permits it to be subsequently elided as a subject – not to mention as an object, as shown in McShane (2005). In (36), the strongly topical antecedent for the elided subject is the nominative case NP *ogromnye zverjugi* 'huge creatures'. *Dolžno byt'* 'it must be' is a fixed, subjectless collocation that functions as a sentence modifier.

(36) – Vot, [e] vidite sledy? Dolžno byt',
Look [e] *see*$_{2.PL.PRES}$ *tracks*$_{ACC.PL}$? *Must*$_{3.SG.NEUT}$ *be*$_{INFIN}$
ogromnye zverjugi! [e] Peresekli
huge$_{NOM.PL}$ *beasts*$_{NOM.PL}$! [e] Crossed$_{3.PL.PAST}$
allejku… (Chm5)
alley$_{ACC.SG.FEM}$
'Look, [you] see the tracks? They must be huge creatures! [They] crossed this path…'

4.2.5. *The Entity's Non-subject Antecedent is, Itself, Elided*

If a category is elided, it must be thematic. For example, *hat* in (37) must be the theme by the end of the first sentence, since in the second sentence it is elided. A reference to the same hat in the next sentence, as a subject, can predictably be elided as well, assuming no intervening clauses that could potentially shift the topic.[10]

(37) Dostal ja šljapu, nadel [e]. Okazalos', [e] nemnožko velilovata, s"ezžaet do nosa, no vse-taki na nej cvety. (Dennis)
'I got the hat and [e] put [it] on. As it turned out, [it] was a little big, [it] slid down to my nose, but there *were* flowers on it.'

To restate this heuristic: if the immediate antecedent of a subject is, itself, elided – no matter its syntactic function – the coreferential subject may be elided as well, given no intervening clauses.

4.2.6. *Elaboration*

Elaboration is a term that has been used with various meanings in the discourse literature. I use it to refer to the second part of a clause complex that is functionally subordinate to the first part but is not joined to it by a conjunction. In writing, the elaboration might be separated by a comma, a semi-colon, a colon, a dash or a period.

One type of elaboration involves restating an event, often by using a more narrow descriptor, adding details, etc. Ellipsis of the subject in the elaborations is predictably permitted and often preferred.

(38) I on slušal, a potom [e] uxodil. [e] Prosačivalsja, kak pesok skvoz' pal'cy. (Tok: 148)
'And he listened, and then [e] left. [He] disappeared like sand between your fingers.'

Detecting a restatement/specification requires semantic analysis. In the OntoSem framework, the ontological relationship between the concepts instantiated by the given verbs can be used to determine how closely they are related and if the second is likely to be a restating or specification of the first. For example, the verbs *leave* and *disappear* in (38) both map to the concept EXIT in the OntoSem lexicon (EXIT is ontologically defined more broadly than the English word 'exit'). Two EXIT events in a row with the same subject referent suggests a restatement, and a restatement permits subject ellipsis.

Another type of elaboration involves introducing a complex event – also known as a script – then providing an enumeration of some of its

[10] The rules of object ellipsis presented in McShane 2005 predict that the 2nd occurrence of *hat* in the first sentence can be elided.

subparts. For example, in (39) the speaker states that the engineer returns home from parents' day, then proceeds to describe what returning home from parents' day is composed of: being at the child's camp then riding home.

(39) «Geroj, inžener tridcati pjati let, letom v voskresen'e vozvraščaetsja s roditel'skogo dnja. [e] Byl u rebenka v lagere. Večerom [e] edet obratno». (Tok: 151)
''The hero, a 35 year-old engineer, is returning one summer evening from parents' day. [He] was visiting his child's camp. In the evening [he] is returning home.''

Similarly, in (40), the narrator states that he started writing a novel then describes what events that was composed of: describing something and trying to convey something else.

(40) Tak ja načal pisat' roman. (Ja) opisal sonnuju v'jugu. [e] Postaralsja izobrazit', kak pobleskivaet pod lampoj s abažurom bok rojalja. (TR)
'And so I started writing a novel. (I) described a sunny snowstorm. [I] tried to convey how the side of a piano shimmers under a shaded lamp.'

In terms of ontological modeling, traveling and writing a novel are both scripts that must be recorded in the ontology available to an intelligent agent. Even very basic versions of the travel and writing-a-novel scripts would suffice to explain the ellipsis in these examples (scripts can become very complex, and need to be for some applications). That is, the most basic description of a travel event is that someone goes somewhere that is not his/her home, is at that place, then comes back – exactly what is needed to understand the acceptability of ellipsis in (39). Similarly, the most basic description of writing a novel includes inventing characters and a plot, describing things, etc. – also perfectly sufficient to explain the acceptability of ellipsis in (40). It must be emphasized that scripts are something that people already know; encoding them in the ontology is simply a formal way of imparting that knowledge to intelligent agents.

The final heuristic for detecting ellipsis-promoting elaboration strategies is the configuration *X does this:* …, or any of its (near-)synonyms or temporal variants.

(41) So svoim sobstvennym proizvedeniem ja postupil tak: [e] uložil ostavšiesja devjat' èkzempljarov i rukopis' v jaščiki stola, [e] zaper ix na ključ i [e] rešil nikogda, nikogda v žizni k nim ne vozvraščat'sja. (TR)
'With my own work I did the following: [I] put the remaning nine copies and the original manuscript in a desk drawer, [e] locked it and [e] decided never ever to return to them again.'

Lexico-semantics, reference resolution (i.e., that *this* corefers with the subsequent text rather than to an antecedent), and punctuation will contribute to the implementation of this heuristic.

4.2.7. *Extensive Previous Description*

Another way that an entity achieves strong topical status is for it to be described at some length prior to the sentence for which subject realization is evaluated, as in (42) and (43).

(42) Na tret'em večere pojavilsja novyj čelovek. Tože literator – s licom zlym i mefistofel'skim, kosoj na levyj glaz, nebrityj. [e] Skazal, čto roman ploxoj, no iz"javil želanie slušat' četvertuju, i poslednjuju, čast'. (TR)
'On the third evening a new person appeared. Also a man of letters – with a mean and devilish face, cross-eyed in his left eye and unshaven. [He] said that the novel was bad but expressed an interest in hearing the fourth and final part.'

(43) – I kak tol'ko moglo prijti v golovu stroit' v takom meste kemping! Net, dlja ètogo malo byt' idiotom, vaš director prosto izvraščenec kakoj-to! – I k tomu že ègoist i proxindej! – s gotovnost'ju podxvatil pan Roman. – Pristan', vidite li, [e] ešče postroil… (Chm5)
''And how could anybody get the bright idea to build a campsite here?! No, for that it's not enough to be an idiot, your boss is some kind of a pervert!'' ''And an egotist and a swindler too!'' Roman added readily. ''And [he] even built a dock …''

Within OntoSem, the extent to which an entity has been described can be detected automatically based on the number of property values recorded for the entity in the text-meaning representations of the immediately preceding context. For example, the text-meaning representation of (43) will include property-value pairs reflecting the meanings of the words *idiot, pervert, egotist* and *swindler* – all of which are applied to the same person (e.g., *idiot* is defined as an ANIMAL whose property INTELLIGENCE has a value of 0 on the abstract scale of {0-1}; similarly, *egotist* is described as a HUMAN whose property MODESTY has a value of 0). It will require further analysis to determine how many descriptors makes an entity sufficiently topical to be elided in a proximate subsequent clause and how other ellipsis promoting and impeding factors interact with this one: for instance, the ellipsis in (43) could be promoted by the genre of dialogue.

4.2.8. *The Ellipsis of 1st and 2nd Person Subjects in Dialogues*

The ellipsis of 1st and 2nd person subjects is widely permitted in Russian, and there is much stylistically neutral optionality. The reason why 1st and 2nd person pronouns can be readily elided stems from the high pragmatic relevance of the interlocutors in any speech act. It is noteworthy that Russian permits subject ellipsis in the past tense as readily as in the present tense even though the past tense – unlike the present tense – does not show person agreement.

(44) [*The speaker first addresses some other people, then the author of the novel*]
– Starik napisal ploxoj, no zanjatnyj roman. V tebe, šel'mec, est' nabljudatel'nost'. I otkuda čto beretsja! Vot [e] už nikak ne ožidal, no!.. soderžanie! (TR)
'This guy wrote a bad but entertaining novel. Man, you have insight. Where do you get this stuff from! [*lit: Where does what come from?*] [I] would never have expected it but!... there's content here!'

(45) Efrosimov. Kogda ja šel po gorodu, èti… nu vot, [e] opjat' zabyl… nu, malen'kie… xodjat v školu?..
'When I was walking around the city these… uh, [I] forgot again… um, the little… [they] go to school?'
Eva. Deti? (Adam)
'Children?'

(46) (*Adam juggles and breaks a glass*)
Anja. Tak. Stakan čužoj! Daraganov stakan.
'Just great. It's not our glass! It's Daragan's.'
Adam. [e] Kuplju stakan. [e] Kuplju Daraganu pjat' stakanov. (Adam)
'[I]'ll buy a glass. [I]'ll buy Daragan five glasses.'

(47) Adam. Segodjna ''Faust'', a zavtra večerom my edem na Zelenyj Mys! Ja sčastliv! Kogda [e] stojal v očeredi za biletami, [e] ves' pokrylsja gorjačim potom i ponjal, čto žizn' prekrasna!.. (Adam)
'Today ''Faust'', and tomorrow evening we're going to Zelyony Mys [*a resort; lit: Green Cape*]. I'm happy! When [I] was standing in line for tickets [I] broke out in a hot sweat and understood that life is wonderful!..'

Unfortunately, it is not the case that *all* 1st and 2nd person subjects can be elided in stylistically neutral Russian dialogues: in some cases, ellipsis adds a colloquial flavor, and in other cases it is simply unacceptable. The

heuristics for making these determinations remain to be discovered. What one *can* do, however, in the near term is compile a list of common locutions that predictably permit subject ellipsis in dialogues, including those presented in (48)–(54). While listing clearly has practical utility, it could also be incorporated into formal language models if such locutions were analyzed as having gained a type of discourse-idiomatic status.

(48) – [e] Ne znaju.
"[I] don't know." [*lit: [e] neg know*$_{1.SG.PRES}$]

(49) – [e] Očen' tebja prošu …
"Please." [*lit: [e] very-much you*$_{ACC}$ *ask*$_{1.SG.PRES}$]

(50) – [e] (Ničego) ne ponimaju.
"[I] don't understand." [*lit: (nothing) neg understand*$_{1.SG.PRES}$]

(51) – [e] [e] Pozdravljaju. (If the subject is overt, so should be the object.)
"Contratulations." [*lit: congratulate*$_{1.SG.PRES.}$]

(52) – Čto [e] budem delat'?
"What are [we] going to do?"

(53) – [e] Predstavljaeš' (sebe)?
"Can [you] imagine?" [*lit: imagine*$_{2.SG.PRES}$ *self*$_{DAT}$]

(54) – [e] Znaete/Znaeš', …
"[You] know$_{\text{FORMAL/INFORMAL}}$, …"

4.3. *Avoidance of Long-Winded Formulations*

In some cases, expressing a subject overtly would require a long-winded description. In such cases, Russian tends to favor eliding the given category (cf. McShane 2005:63-65 for comparable cases of object ellipsis in Russian). One such case, illustrated by (55), involves indefinite referents: the meaning of the subject is *that person, whoever he or she may be*.

(55) [*Our heroine wants someone else to call her boyfriend's hotel room to see if he's there. She's thinking…*]
Možet, kto-nibud' drugoj za menja pozvonit? Kto ugodno, liš' by ne ja. [e] Poprosit soedinit' i proverit, na meste li on. (Chm6: 16)
'Maybe somebody else could call for me? Anyone, as long as it's not me. [They/That person] will ask to be connected and will check if he's there or not.'

In colloquial English, we tend to use *they* in such contexts to avoid the he/she problem, despite the blatant lack of number agreement this work-around entails.

Another situation in which subject ellipsis is conveniently preferred is when the referent is not an object but a (sometimes vaguely delineated) situation, as shown by examples (56) and (57).

(56) [*A woman's friend has just said that she thought she heard, over the phone, that the first woman's boyfriend was with another woman. The first woman responds*:]
[e] Očen' na nego poxože. (Chm6: 21)
'[That] sounds a lot like him. [*lit: [e] very-much to him similar*]'

(57) Snačala my peli vse xorom. ''Vo pole berezon'ka stojala''.
[e] Vyxodilo očen' krasivo… (Dennis)
First we all sang ''There was a birch tree in the forest''.
'[It] came out really well.'

In (56), the elided subject means something like *the fact that he is with another woman*. If the subject were overt, it would need to be *èto*, 'that', or some descriptive locution. In (57), the elided subject refers to the singing of the song, not the song itself (*it* in English is ambiguous between the two interpretations); here, neither *èto* 'it/that' nor anything else that is reasonably terse could be used overtly in Russian.

The prerequisite for incorporating a rule for convenient ellipsis into the language processor of an intelligent agent is a very sophisticated level of semantic analysis and reference resolution. In a context like (56), it is very difficult to automatically establish what span of the preceding context is being referred to; only if this is done successfully can a system determine whether the meaning of that span suggests the use of convenient ellipsis in Russian. Most natural language processing systems do not touch this aspect of reference resolution, constraining work to much simpler contexts in which nominal referring expressions have nominal antecedents.[11]

5. Discussion

This paper has developed an approach to the analysis of subject ellipsis in Polish and Russian that, in terms of genre, borrows more from the field of artificial intelligence than from current mainstream linguistics. Although the generalizations are not sequentially placed in an algorithm or written in a particular metalanguage for a particular language processing system, they could be, as the heuristics are sufficiently formal to be machine tractable. Machine tractability is one litmus test for the

[11] Work by Donna Byron (2004) and colleagues is one exception. Within OntoSem we are working on this problem within an all-purpose reference resolution module.

degree of formalization of an approach: if a machine can compute it, it must be formal. In this way, I counter the typical objection to approaches that embrace semantics and pragmatics: not formal enough.

This model cannot yet confidently predict every subject realization choice in Russian and Polish but it takes a step in that direction. In the next stage of the work, the model must additionally incorporate dependencies in ellipsis, as illustrated by the following two Russian examples. In (58) if both arguments are not elided then both must be overt. Having one overt but not the other would be very strange (the point at which strangeness becomes ungrammaticality is an open question).

(58) Miška sel so mnoj i vzjal v ruki samosval.
- Ogo! – skazal Miška. – Gde [e] dostal [e]? (Dennis)
– where [e] get [e]?
Mishka sat down with me and picked up my dump truck.
"Wow!," said Miskha. "Where'd you get it?"

Likewise, in (59) the fact that the underlined instances of *ja* are overt means that the respective main-clause subjects must be elided; not eliding them would be stylistically infelicitous.

(59) Dal'še razmylo v pamjati mesjaca dva. [e] Pomnju tol'ko, čto ja u Rudol'fi vozmuščalsja tem, čto on poslal menja k takomu, kak Rvackij, čto ne možet byt' izdatel' s mutnymi glazami i rubinovoj bulavkoj. [e] Pomnju takže, kak eknulo moe serdce, kogda Rudol'fi skazal: "A pokažite-ka vekselja", – i kak ono stalo na mesto, kogda on skazal skvoz' zuby: "Vse v porjadke". Krome togo, [e] nikogda ne zabudu, kak ja priexal polučat' po pervomu iz ètix vekselej. (TR)
'The next two months are a blur. [I] remember only that I was outraged at Rudol'fi for sending me to a guy a like Rvatskij, who couldn't be a publisher with those muddy eyes and that ruby-studded pin. [I] remember also how my heart skipped a beat when Rudol'fi said, "Show me the IOU," and how it relaxed when he muttered, "Everything's alright." Apart from that [I] will never forget how I went to pick up the payment from the first of those IOUs.'

It is not clear yet if fine stylistic subtleties can be formally modeled, but there is no reason yet to believe they cannot.

The analysis presented here contributes to the theory of ellipsis developed in McShane 2005, whose goal is the development of an inventory of parameters and values that is sufficient to describe and formally model ellipsis cross-linguistically. Although the heuristics in this paper were not formulated as parameters and values, such recasting is only a matter of form, not content, as shown by the examples in Table 4.

Table 4. Recasting heuristics as parameters and values

Parameter	Values
Basic pronominal subject status	Elided / overt
Syntactic configurations requiring an overt subject in Polish	The heuristics in the subsections of 3.1
Sources of potential referential ambiguity for subjects	The heuristics in the subsections of 3.2
…	…

The inventory of parameters and values developed for one language should help to guide work in other languages, with iterative refinement of the inventory a necessary and welcome outcome.

The work presented here should be useful "off the shelf" to computational linguists, descriptive grammarians, and anyone interested in usage rules of subject ellipsis in Russian and Polish. The implications for syntactic theory remain to be formulated, but it is hoped that this data and analysis can serve as a good starting point for researchers in generative syntax as well as a stimulus to carry out rigorous corpus analysis for other languages. It is clear that, as so-called interface phenomena gain wider attention, multi-modular analyses will become ever more common.

The success of the current approach will ultimately be judged by the ability of a non-native speaker or intelligent agent (a) to successfully predict when subject ellipsis should and should not be used when generating utterances that express a specific meaning in a specific context and (b) to detect when a subject has been elided and determine whether or not the ellipsis decision carries semantic or pragmatic implications.

In contrast to much theoretical work, no claims have been made about "explaining" component phenomena, as that term has been overused in the field of linguistics, with its current semantics being as light as "corresponding with the model set forth." While the definition of explanation will be left to the philosophers of science, the current work claims only to present a useful, linguistically grounded, machine-tractable model that accounts for real-world data in a compact way and suggests the types of features and contexts that should be considered when studying subject ellipsis cross-linguistically.

Sources for citations (abbreviation then full listing)

Adam BULGAKOV, M. *Adam i Eva*. (http://www.lib.ru; in Russian)
Chm1 CHMIELEWSKA, J. *2/3 sukcesu*. (http://www.chmielewska.art.pl/bibliografia; in Polish)
Chm2 CHMIELEWSKA, J. *Jeden kierunek ruchu*. (ibid)
Chm3 CHMIELEWSKA, J. *Upiorny legat*. (ibid)
Chm4 CHMIELEWSKA, J. *Wielki diament*. (ibid)

Chm5 KHMELEVSKAJA, I. *Osobye zaslugi*. (http://www.aldebaran.com.ru; in Russian)
Chm6 KHMELEVSKAJA, I. 1997. *Klin klinom*. Polish to Russian translation by L. Ermilova. Moscow: Phantom Press International.
Dennis DRAGUNSKIJ, V. *Deniskiny rasskazy*. (http://www.lib.ru; in Russian)
IPI The IPI PAN (Polish) corpus, at http://korpus.pl/
Pok HERBERT, Z. 1976. Drugi pokój. *Antologia Dramatu*. Warszawa: Państwowy Instytut Wydawniczy. (In Polish)
Tok TOKAREVA, V. 1995. *Korrida*. Moscow: Vagrius. (In Russian.)
TR BULGAKOV, M. *Teatral'nyj roman*. (http://www.lib.ru; In Russian.)

References

AVRUTIN, S. & BABYONYSHEV, M. 1997. Obviation in subjunctive clauses and Agr: Evidence from Russian. *Natural Language and Linguistic Theory* 15, 229–262.
BYRON, D. K. 2004. Resolving pronominal reference to abstract entities. Ph.D. Dissertation & Technical Report 815, University of Rochester Computer Science Department.
FRANKS, S. 1995. *Parameters of Slavic Morphosyntax*. Oxford: Oxford University Press.
GORDISHEVSKY, G. & AVRUTIN, S. 2004. Optional omissions in an optionally null subject language. *Proceedings of GALA 2003, vol. 1*, ed. J. van Kampen & S. Baauw, 187–198.
GRIMSHAW, J. & SAMEK-LODOVICI, V. 1998. Optimal subjects and subject universals. *Is the Best Good Enough: Optimality and Competition in Syntax*, ed. P. Barbosa, D. Fox, P. Hagstrom, M. McGinnis & D. Pesetsky, 193–219. Cambridge, MA: MIT Press.
LINDSETH, M. 1998. *Null-Subject Properties of Slavic Languages (with special reference to Russian, Czech and Sorbian)*. München: Verlag Otto Sagner.
LOBECK, A. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*. New York; Oxford: Oxford University Press.
MCSHANE, M. 1998. *Ellipsis in Slavic: The Syntax-Discourse Interface*. Unpublished Ph.D. Dissertation, Princeton University.
MCSHANE, M. 1999a. Predictive rules of direct object ellipsis in Russian. *Formal Approaches to Slavic Linguistics, The Seattle Meeting 1998*, ed. K. Dziwirek, H. Coats & C. M. Vakareliyska, 329–348. Ann Arbor, Michigan: Michigan Slavic Publications.
MCSHANE, M. 1999b. The ellipsis of accusative direct objects in Russian, Polish and Czech. *Journal of Slavic Linguistics* 7, 45–88.
MCSHANE, M. 2000a. Verbal ellipsis in Russian, Polish and Czech. *Slavic and East European Journal* 44, 195–233.
MCSHANE, M. 2000b. Hierarchies of parallelism in elliptical Polish structures. *Journal of Slavic Linguistics* 8:, 83–117.
MCSHANE, M. 2005. *A Theory of Ellipsis*. Oxford; New York: Oxford University Press.
MCSHANE, M., NIRENBURG, S. & BEALE, S. 2005a. Semantics-based resolution of fragments and underspecified structures. *Traitement Automatique des Langues* 46, 163–184.

McShane, M., Nirenburg, S. & Beale, S. 2005b. An NLP lexicon as a largely language independent resource. *Machine Translation* 19, 139–173.

McShane, M., Beale, S. & Nirenburg, S. 2004. OntoSem methods for processing semantic ellipsis. *Proceedings of the HLT/NAACL 2004 Workshop on Computational Lexical Semantics*.

McShane, M., Nirenburg, S., Beale, S., Jarrell, B. & Fantry, G. 2007. Knowledge-based modeling and simulation of diseases with highly differentiated clinical manifestations. *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 07)*, Amsterdam, The Netherlands, July 7–11, 2007.

McShane, M., Jarrell, B., Fantry, G., Nirenburg, S., Beale, S. & Johnson, B. 2008. Revealing the conceptual substrate of biomedical cognitive models to the wider community. *Medicine Meets Virtual Reality* 16, ed. J. D. Westwood, R. S. Haluck, H. M. Hoffman, G. T. Mogel, R. Phillips, R. A. Robb, K. G. Vosburgh, 281–286.

Merchant, J. 2004. Fragments and ellipsis. *Linguistics and Philosophy* 27, 661–738.

Mitkov, R. 2001. Outstanding issues in anaphora resolution. *Computational Linguistics and Intelligent Text Processing*, ed. A. Gelbukh, 110–125. Springer.

Nilsson, B. 1982. *Personal Pronouns in Russian and Polish: A study of their communicative function and placement in the sentence* (translated from Swedish by Charles Rougle). Stockholm-Sweden: Almqvist & Wiksell International.

Nirenburg, S. & Raskin, V. 2004. *Ontological Semantics*. Cambridge, MA: MIT Press.

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R. & Stevenson, R. 2000. Specifying the parameters of Centering Theory: a corpus-based evaluation using text from application-oriented domains. *Proceedings of the 38th ACL*, Hong Kong, October.

Stainton, R. 2006. Neither fragments nor ellipsis. *The Syntax of Nonsententials*, ed. L. Progovac, K. Paesani, E. Casielles & E. Barton, 93–116. Philadelphia: John Benjamins.

Traum, D. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Thesis and TR 545, Computer Science Dept., U. Rochester, December 1994.

Marjorie McShane
Department of Computer Science and Electrical Engineering
Institute for Language and Information Technologies
ITE 325
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
USA
marge@umbc.edu