Zleceniodawca: IPI PAN **Dzieło do umowy z dnia: 2.01.2017**
Zleceniobiorca: Jakub Kozakoszczak **Termin zakończenia pracy: 31.01.2016**

Representation of opacity and quantification in a semantic parser for Polish

Wojciech Jaworski **Jakub Kozakoszczak**
Institute of Computer Science, Polish Academy of Sciences
University of Warsaw
wjaworski@mimuw.edu.pl jkozakoszczak@gmail.com

Zrealizowane w ramach projektu:
„CLARIN – Polskie wspólne zasoby językowe i infrastruktura technologiczna"
Tytuł pracy zamówionej:
„Integrated representation of semantic phenomena beyond events and roles for implementational purposes"
Adres dzieła:
http://wiki.nlp.ipipan.waw.pl/clarin/Parser%20kategorialny?action=AttachFile&do=view&target=JK_phenomena_beyond_events.zip
Opracowanie dokumentu: Jakub Kozakoszczak

Abstract

In this paper we present a semantic metalanguage which integrates a standard neo-Davidsonian approach with representation of the phenomenon of opacity and MRS-style underspecification of the scope of quantifiers.

1 Introduction

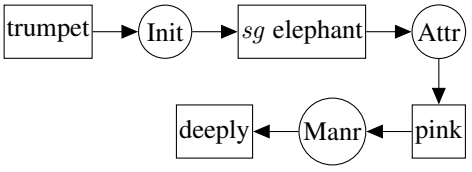
ENIAM is a deep semantic parser for sentences in Polish. The semantic metalanguage it employs is based on the standard neo-Davidsonian approach in that it reifies all events and expresses the role of their participants in terms of thematic roles. Due to the use of *nested contexts* ENIAM's metalanguage is also capable of expressing more subtle phenomena connected to opacity. The intended utility of the metalanguage is to support all the language processing tasks that involve the semantic level, in particular Information Retrieval, Question Answering and Recognizing Textual Entailment but first and foremost the system is designed for research use in the humanities and social sciences dealing with large collections of Polish texts.

2 Broad reification

In ENIAM's metalanguage almost all concepts are consistently reified. Every lexeme (or MWE) which is not a quantifier, conjunct or non-semantic item is translated into a variable identifying the entities under discussion and a dyadic TYPE or HASNAME predicate with a constant term identifying the lexeme's intension. The main reason for the broad reification is the modifiability of virtually every part of speech in Polish, including adjectives, adverbs and propositions, see adverb + adjective "intensywnie różowy" 'in deep pink' below.

(1) *Intensywnie* *różowy* *słoń* *trąbi.*
Deeply pink elephant trumpets.
'An elephant in deep pink trumpets.'

$$\begin{aligned}&\exists(t, \text{TYPE}(t, \text{trumpet}),\\ &\exists(e, \text{TYPE}(e, \text{elephant}) \wedge |e| = 1,\\ &\exists(p, \text{TYPE}(p, \text{pink}), \exists(d, \text{TYPE}(d, \text{deeply}),\\ &\text{INITIATOR}(t, e) \wedge \text{ATTR}(e, p) \wedge\\ &\text{MANNER}(p, d)))))\end{aligned} \quad (2)$$



(3)

3 Lexical and grammatical meanings

The semantic representation is built upon dependency trees augmented with concepts (word senses) and with thematic roles. Concepts are ascribed to lexemes and originate from Słowosieć (the Polish WordNet) (Maziarz et al., 2014). Each sense is represented in the WordNet style as a lemma with a number. The senses of proper names are their types.

The valency of the lexemes in the sentence is determined with the valency dictionary Walenty (Przepiórkowski et al., 2014). Walenty covers most verbs and many nouns, adjectives and adverbs. Each entry comprises syntactic schemata that include detailed syntactic description of the obligatory dependents and, importantly, semantic frames that give their semantic characteristic, namely thematic roles and selectional preferences. The schemata and the frames are mapped many-to-many. The set of thematic roles employed in ENIAM's metalanguage extends the set of thematic roles provided by Walenty.

4 Semantic Graphs and quantifiers' scope underspecification

The logical formulae are presented in the form of semantic graphs (SG) inspired by Sowa's conceptual graphs (Sowa, 1984).

The boxes represent entities mentioned in the text. The first one represents the action of *trumpeting*, the second one represents the *elephant* (a singleton of elephant-type entities) The symbol *sg* is a quantifier that defines the count of the elephants under discussion as exactly one. The circles represent relations between (or thematic roles of) the entities. The Init relation says that the *elephant* is the initiator of *trumpeting*.

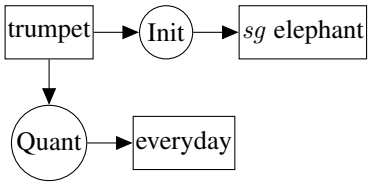
Since the example (1) is unambiguous (spuriously ambiguous) with regard to quantifier scopes, its semantic graph (3) which is the actual[1] representation in ENIAM's metalanguage is equivalent to the single logical form given in (2). However, opposite to FOL, SGs underspecify the scope of the quantifiers. SGs are equivalent to weakly specified MRS structures (Copestake et al., 2005) where all restriction holes are resolved but none of the body holes – which in practice boils down to no constraints of equality modulo quantifiers being defined. Since SGs are special cases of MRS structures, each SG defines a set of standard FOL formulae that are the allowed readings of the parsed sentence.

[1]For the convenience of non-Polish speaking readers all presented logical formulae are translated into English and are therefore not identical to the parser output. In particular the parser presents concept names in Polish.

5 Natural language quantifiers

We extend FOL with special quantifiers existing in the language, e.g. *co dziesiąty* ('every tenth'), *prawie każdy* ('almost every') or *codziennie* ('every day')

(4) *Słoń* *codziennie* *trąbi.*
Elephant everyday trumpets.
'An elephant trumpets everyday.'



(5)

One of the possible readings defined by SG (5) is

$$\exists(s, \text{TYPE}(s, \text{elephant}) \wedge |s| = 1, \\ \text{EVERYDAY}(t, \text{TYPE}(t, \text{trumpet}) \\ \wedge \text{INITIATOR}(t, s))) \quad (6)$$

6 Nested contexts

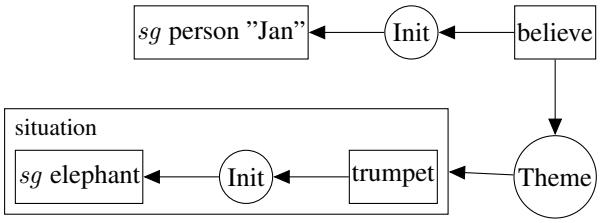
For the majority of natural language sentential operators their meaning is not extensional, so their sentential operands are in opaque contexts

(7) *Jan* *wierzy,* *że* *słoń* *trąbi.*
Jan believes, that elephant trumpets.
'Jan believes that an elephant trumpets.'

In the sentence above, the object of belief of Jan has the role of a belief's Theme. In order to express the fact we make use of *nested contexts* (Sowa, 1984, p. 173). We assume an extended version of FOL with one dyadic meta-predicate DSCR that binds a formula to its identifier, see $\text{DSCR}(b, \dots)$ below. The identifier b is allowed as an argument of a predicate, in particular a thematic role THEME.

$$\exists(w, \text{TYPE}(w, \text{believe}) \wedge \exists(j, \text{TYPE}(j, \text{person}) \wedge \\ \text{HASNAME}(j, \text{'Jan'}) \wedge |j| = 1, \\ \text{INITIATOR}(w, j)) \wedge \\ \exists(b, \text{TYPE}(b, \text{situation}) \wedge \\ \text{DSCR}(\ b, \exists(s, \text{TYPE}(s, \text{elephant}) \wedge |s| = 1, \\ \exists(t, \text{TYPE}(t, \text{trumpet}) \wedge \text{INITIATOR}(t, s)))\ ), \\ \text{THEME}(w, b))) \quad (8)$$

In SC we nest an inner SC representing the sentential complement in an additional box of the type (labeled as) "situation".



(9)

7 Inner models

The propositions represented in a nested context are not implied by the whole sentence and they describe a separate inner model. The meaning of nested contexts is to indicate the inner models.

The semantics of the metalanguage is build upon a set of possible worlds. We assume that each speaker has their own world model in which they interpret their utterances and each act of communication creates of a new model. Obviously, when someone communicates something, it doesn't mean they believe it, but they present a model in which it is true. Thanks to representing each utterance as placed in a separate context the very representation of a longer text which includes many contradictory opinions isn't contradictory itself.

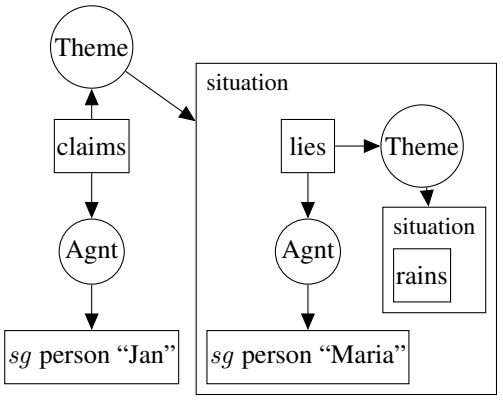
Some contexts that introduce a new model include:

1. imperatives
2. conditionals
3. speech verbs (*say*, *claim*, *confirm*)
4. epistemic (modal) verbs (*believe*, *assume*)
5. modal adverbs and particles (*probably*, *presumably*)

8 Reported speech

In the case when someone refers what another person said or believes the former simulates in their own model the model of the latter. This is represented as nesting of the SC for the second model in a context within the SC for the first model. The nesting goes as deep as needed.

(10) *Jan twierdzi, że Maria kłamie, że pada.*
Jan claims that Maria lies that rains
'Jan claims that Maria lies that it rains.'



(11)

Representation (11) can be understood as:

1. The narrative introduces a model $\mathcal{A}$ in which a situation a takes place, namely *Jan claims that b*.
2. $\mathcal{B}$ is a model in which the situation b takes place, namely *Maria lies that c*.
3. $\mathcal{C}$ is a model in which the situation c takes place, namely *it rains*.

Or more formally:

$$\begin{array}{l}\text{The speaker communicates that a}\\ \wedge\ \textsc{dscr}(a,\ \text{Jan claims that b}\\ \wedge\ \textsc{dscr}(b,\ \text{Maria lies that c}\\ \wedge\ \textsc{dscr}(c,\ \text{It rains})))\end{array} \quad (12)$$

If Jan's claim is false there is a discrepancy between the model $\mathcal{B}$ and the actual world. If Maria's statement is false there is a discrepancy between the model $\mathcal{C}$ and the actual world.

9 Factivity

In ENIAM's metalanguage factivity is not taken to be a property of the matrix verb (like *to know that*) but against tradition – of a particular syntactic position of the verb. The reason for this is the fact that some Polish verbs can simultaneously take multiple propositional complements of different syntactic types introduced by different complementizers. Such complements can be coordinated yet differ with respect to factivity. In the example below the verb *nauczyć się* 'to learn' takes one complement introduced by *że* to give the meaning 'to learn that' conjoined with another one introduced by *żeby* to the meaning 'to learn to'. The truth of the first type of complement is implied by the embedding sentence but the second never is.

(13) *Nauczyłem się, że w mieszkaniu*
Learned.1SG REFL ŻE in flat
ważne jest dobre oświetlenie i
important COP good lightning and
żeby nie czytać po ciemku.
ŻEBY not read.INF dark-like
'I learned that good lightning is important in a flat and [I learned] to not read in the dark.'

At the same time non-factivity is not a persistent property of *żeby*-complements. In the next example the *żeby*-complement of the verb *wymusić* 'to force into' is in a factive position.

(14) *Wymusiła na mnie, żebym*
Forced.3SG.F.PST on me ŻEBY.1SG
porobił zdjęcia.
make photographs
'She forced me into taking some photographs.'

The representation of a factive complement is not different from other complements, so it is nested in an inner context. Special treatment of the information the complement bears is due to the lexical information about the factivity of this syntactic position of this lexeme stored and due to the **axiom of factivity**. The axiom says that if a subordinate clause is in a factive position then if the logical form of the superordinate clause is true in a model then the logical form of the subordinate clause is also true in the model.

Axiom of factivity Let $\text{FACTIV}(\varphi, \psi)$ signify that ψ represents a subordinate clause in a factive syntactic position and φ represents the superordinate clause. Then

$$\begin{gathered}\text{For each model } \mathcal{M} \text{ and formulae } \varphi \text{ and } \psi, \\ \text{if } \text{FACTIV}(\varphi, \psi) \\ \text{then } (\mathcal{M} \vDash \varphi \longrightarrow \mathcal{M} \vDash \psi))\end{gathered} \tag{15}$$

10 Resources

Apart of two large semantic resources – Słowosieć and Walenty – which were created as a part of CLARIN-PL project, the system ENIAM benefits from own lexical resources comprising particles, adjectives, adverbs and PPs with opaque or quantificational meaning, and factive valency positions of verbs.

Acknowledgements

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

References

A. Copestake, D. Flickinger, C. Pollard, and I.A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language & Computation*, 3(4):281–332.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, and Stan Szpakowicz. 2014. plwordnet as the cornerstone of a toolkit of lexico-semantic resources. In *Proceedings of the Seventh Global Wordnet Conference*, pages 304–312.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

J. F. Sowa. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.