Syntactic Approximation of Semantic Roles

Wojciech Jaworski[1,2] and Adam Przepiórkowski[1]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland,
adamp@ipipan.waw.pl
[2] Institute of Informatics, University of Warsaw,
ul. Banacha 2, 02-097 Warsaw, Poland,
wjaworski@mimuw.edu.pl

Abstract The aim of this paper is to propose a method of simulating – in a syntactico-semantic parser – the behaviour of semantic roles in case of a language that has no resources such as VerbNet of FrameNet, but has relatively rich morphosyntax (here: Polish). We argue that using an approximation of semantic roles derived from syntactic (grammatical functions) and morphosyntactic (grammatical cases) features of arguments may be beneficial for applications such as text entailment.

Key words: thematic roles, parser, morphosyntax, LFG

1 Introduction

There is a strong tradition in Slavic linguistics of relating morphosyntax to semantics, especially, of claiming that morphological cases have unified meanings. One of the most prominent proponents of this approach was Roman Jakobson (see, e.g., Jakobson 1971a,b), and it has been further developed by Anna Wierzbicka (e.g., Wierzbicka 1980, 1981, 1983, 1986), who claims that "cases have meanings and that this meaning can be stated in a precise and illuminating way" (Wierzbicka, 1986, p. 386).

While we do not fully subscribe to this tradition, we show that it turns out to be a useful approach in Natural Language Processing (NLP). In particular, we discuss the role of semantic roles in grammar engineering and argue that – in case of languages with rich morphosyntax but no manually created semantic role resources such as VerbNet or FrameNet – a relatively simple way of inferring an approximation of semantic roles from syntax and morphosyntax may be sufficient for some applications. In fact, it seems that even when a resource like VerbNet is available, this simpler approach to semantic-like roles may be beneficial.

The broad aim of the work partially reported here is to add a semantic component to the manually created LFG (Bresnan, 2001; Dalrymple, 2001) grammar of Polish (Patejuk and Przepiórkowski, 2012), implemented using the XLE platform (Crouch *et al.*, 2011). Regardless of this particular context, we believe that the approach proposed in Sect. 3 has a wider applicability.

2 Semantic roles in grammar engineering

The modern notion of semantic roles stems from the work of Gruber 1965 and Fillmore 1968, and it was brought to the foreground of linguistic research by Jackendoff 1972. Relatively small sets of semantic roles are commonly assumed in theoretical linguistics. For example, Fillmore 1968 distinguishes between Agentive, Dative, Instrumental, Factive, Locative, Objective, as well as Benefactive, Time and Comitative, and even fewer roles are assumed in LFG (Bresnan and Kanerva, 1989; Dalrymple, 2001). In more applicational or corpus-based work, much larger repertoires are adopted, e.g., 18 roles in the system of Sowa 2000 or 30 roles in VerbNet (Kipper *et al.* 2000; `http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`).

Semantic roles are useful in those NLP tasks which use or produce semantic representations for the purpose of automatic reasoning, e.g., in text entailment or question answering. For example, instead of representing the sentence *Carrie ate pizza at Langley* naïvely as $\exists p\, \mathit{pizza}(p) \wedge \mathit{eat}(C, p, L)$, it may be a little less naïvely (but still ignoring tense, etc.) represented using semantic roles and the neo-Davidsonian approach (Parsons, 1990) as $\exists p, e\, \mathit{pizza}(p) \wedge \mathit{eat}(e) \wedge \mathit{agent}(e, C) \wedge \mathit{patient}(e, p) \wedge \mathit{location}(e, L)$. This latter representation makes the inference from *Saul ate pizza at Langley* to *Saul ate pizza*, represented as $\exists p, e\, \mathit{pizza}(p) \wedge \mathit{eat}(e) \wedge \mathit{agent}(e, C) \wedge \mathit{patient}(e, p)$, immediate – it's a matter of dropping the conjunct $\mathit{location}(e, L)$ in the semantic representation. On the other hand, on the more traditional approach, many meaning postulates would have to be formulated, including one relating the 3-argument *eat* predicate (as in $\mathit{eat}(C, p, L)$) to the corresponding 2-argument predicate (as in $\mathit{eat}(C, p)$).

Given the multiplicity of proposed systems of semantic roles, the question arises which one to use in a grammar engineering task. In Jaworski and Przepiórkowski 2014 we report the results of usability studies of the two systems mentioned above: Sowa's and VerbNet. The results are discouraging: the inter-annotator agreement is much too low to guarantee a reasonable quality of semantic role assignment – and, hence, the quality of any tools trained on corpora annotated with such semantic roles – and the investigation of disagreements reveals some internal inconsistencies in these systems.

On the basis of these experiments, as well as various remarks in the literature, we conclude that semantic role systems such as VerbNet or Sowa's are not really well-suited for the grammar engineering task and that other approaches must be explored. The one that we advocate here is to define 'semantic roles' on the basis of morphosyntactic information, including morphological cases, following the linguistic tradition referred to at the beginning of this section. This tradition is continued by Slavic linguists working within the Cognitive Linguistics paradigm, including Ewa Dąbrowska, whose view of the Polish dative reads like a definition of a semantic role: "the dative noun refers to an individual affected by a process or state which obtains in some part of his personal sphere, be it the sphere of potency, the sphere of empathy, the sphere of awareness, or the private sphere" (Dąbrowska 1997, p. 68; see also Dąbrowska 1994).[3]

3 Syntactic approximation of semantic roles

There are two general approaches to obtaining semantic representations in LFG-based parsing systems: co-description (CD) and description-by-analysis (DBA). The former, CD, is straightforward: lexical entries contain lexical semantic information, grammar rules or principles specify how meanings are composed, and semantic composition proceeds in parallel to syntactic parsing. This is the standard procedure in various formalisms and parsing platforms, including the HPSG (Pollard and Sag, 1994) English Resource Grammar (Copestake and Flickinger 2000; `http://www.delph-in.net/erg/`).

However, in LFG grammar engineering, the second approach, DBA, is common: semantic representation is obtained by analysing f-structures, i.e., non-tree-configurational syntactic representations (as opposed to more surfacey tree-configurational c-structures) containing information about predicates, grammatical functions and morphosyntactic features; this approach has been adopted for German (Frank and Erk, 2004; Frank and Semecký, 2004; Frank, 2004), English (Crouch and King, 2006) and Japanese (Umemoto, 2006).

In order to obtain representations employing semantic roles, resources external to the respective LFG grammars must be used in the process. Thus, in case of German, rules of transforming f-structures to semantic structures containing semantic role information were automatically acquired (Frank and Semecký, 2004) on the basis of a German treebank (Brants *et al.*, 2002) annotated with FrameNet-like information, and subsequently generalised (Frank, 2004) to cover more unseen cases. For English, semantic roles were more directly transferred from VerbNet to the lexicon (Crouch and King, 2005) used in the system rewriting f-structures to semantic representations (Crouch and King, 2006). On the other hand, apparently no such external resources were used in case of Japanase (Umemoto, 2006), so the resulting representations use the names of grammatical functions such as subject and object, instead of true semantic roles.

Frank and Erk 2004 point out the benefits of adopting the DBA approach, especially at an early stage of developing a semantic module of an LFG parser, and we follow this advice here. However, there are currently no external resources for Polish that could supply information about semantic roles of particular predicates. But instead of falling back all the way to grammatical functions, as in case of the Japanese parser mentioned above, we capitalise on the fact that Polish has a relatively rich morphosyntactic system, with 7 morphological cases,[4]

[3] In this approach, Polish and English have different sets of semantic roles built into their morphosyntactic structure. The interesting philosophical impliciation of this point of view – very much in the spirit of the so-called Sapir-Whorf hypothesis – is that users of Polish and English perceive the world differently, i.e., they have different categorisation of relations between events and their participants.

[4] But only 6 of them are governable; the vocative is never governed.

and a large number of preposition / morphological case combinations, many of which are highly correlated with specific semantic roles. In the remainder of this section we describe the procedure of assigning 'semantic roles' on the basis of morphosyntactic information; to constantly remind ourselves that they are just approximations of true semantic roles, they will be called R0, R1, etc., instead of Agent, Patient, etc., and the term 'semantic role' will be written in scare quotes.

How many roles do we need? We have seen above that too many roles cause classification problems, so we want as few different roles as possible. On the other hand, there should be enough of them – and they should be sufficiently well differentiated – to minimise the probability of two arguments of the same predicate bearing the same role.[5] For the time being we settle on 11 core roles listed in Tab. 1 together with the meanings they are supposed to approximate (and, in some cases, with the usual names of such roles).

Role	**Approximate description**
R0	Actor of an action (Agent, Effector)
R1	Undergoer of an action (Patient, Theme, Product)
R2	Dative argument (Beneficiary, Recipient)
R3	Instrumental argument (Instrument)
R4	Adlative argument in both physical and abstract (functional, purposive) meaning (Destination, Recipient, Theme)
R5	Ablative argument in both physical and abstract (causal) meaning (Source)
R6	Locative argument in both physical and abstract meaning
R7	Perlative argument
R8	Topic of communication
R9	Temporal argument (point in time)
R10	Manner argument

Table 1. Assumed 'semantic roles' and their approximate meanings

The algorithm for assigning 'semantic roles' to arguments is rather simple. With one exception, the 'semantic role' is assigned on the basis of the grammatical function of the argument (as well as the voice of the verb; see below). The exception is the OBL(ique) argument – in the LFG grammar of Polish this is prototypically the grammatical function of various prepositional arguments. In this case, also the form of the preposition and the case of its object is taken into account. Tables 2 and 3 present the mapping from grammatical functions of arguments of an active form of the verb, and – in case of OBL – from particular preposition / case combinations, to 'semantic roles'.[6] In case of passive forms,

[5] Note that it would be unrealistic to expect such situation never to happen; even VerbNet with its 25–30 roles needs roles such as Co-Agent, Co-Patient and Co-Theme. Moreover, in the experiments described in Jaworski and Przepiórkowski 2014, about 2.5–4.4% of verb occurrences had their arguments marked with duplicated roles (more precisely, 2.47% in case of VerbNet roles, 4.36% in case of Sowa's roles).

[6] The mapping given for OBL may also be used to assign 'semantic roles' to prepositionial adjuncts. Roles R6, R9 and R10 may also be used to indicate relations between

Argument	**Role**
SUBJ	R0
OBJ	R1
OBJ-TH	R2
OBL-INST	R3
OBL-GEN	R1
OBL-STR	R1
OBL	see Tab. 3
XCOMP	R8
COMP	R8
XCOMP-PRED	R8

Table 2. Mapping of grammatical functions (with active verbs) to 'semantic roles'

Preposition / morphological case	**Role**
DLA[gen], PRZECIW[dat], WOBEC[gen]	R2
DO[gen], KU[dat], MIĘDZY[acc], NA[acc], NAD[acc], PO[acc], POD[acc], POMIĘDZY[acc], PONAD[acc], POZA[acc], PRZED[acc], W[acc], ZA[acc]	R4
DZIĘKI[dat], OD[gen], SPOD[gen], SPOŚRÓD[gen], WSKUTEK[gen], Z[gen], ZZA[gen]	R5
KOŁO[gen], MIĘDZY[inst], NA[loc], NAD[inst], PO[loc], POD[inst], POMIĘDZY[inst], PONAD[inst], PONIŻEJ[gen], POZA[loc], PRZED[inst], PRZY[loc], U[gen], W[loc], WOKÓŁ[gen], WŚRÓD[gen], ZA[inst]	R6
BEZ[gen], POPRZEZ[acc], PRZEZ[acc], Z[inst]	R7
JAKO[nom], O[acc], O[loc]	R8
PODCZAS[gen]	R9
WEDŁUG[gen]	R10

Table 3. 'Semantic roles' for OBL arguments

the deep object becomes the surface subject, so SUBJ maps to R1, and – conversely – the deep subject may be realised as a *by*-phrase (PRZEZ[acc]) bearing the OBL-AG grammatical function, so this function is mapped to R0.

This way of assigning 'semantic roles' conflates different grammatical functions while preserving the near-uniqueness of 'semantic roles'. First, normally only one of the grammatical functions OBJ (any passivisable argument, usually in the accusative), OBL-GEN (non-passivisable genitive argument) and OBL-STR (structurally cased, i.e., a usually accusative argument, which does not passivise) may appear in the f-structure of a given verb, so these – as well as the SUBJ of a passive verb – are uniformly mapped to R1. Second, in the valence dictionary for Polish mentioned below, there is only one rather special verb that has a valence schema with different arguments mapping to the grammatical functions of COMP (sentential complement) and XCOMP (infinitival complement), so it makes sense to translate both grammatical functions to R8, which approx-

a verb and its adverbial adjuncts. But if adjuncts are included in the 'semantic role' assignment, the problem of duplication of roles mentioned below becomes more serious and should be dealt with, e.g., by assigning adjuncts separate roles A2–A10, conventionally corresponding to R2–R10.

imates the Topic role. Less obviously, also XCOMP-PRED, which corresponds to the predicative argument of copula verbs, especially, BYĆ 'be' and ZOSTAĆ 'become', is translated to R8. It might at first seem that a sentence meaning *Brody is innocent* should be represented as, say, *innocent*(b), but then there is no event that different tenses or modalities could modify. Without going into details of the envisaged semantic representation, let us assume that a proposition like *innocent*(b) – expressed by the XCOMP-PRED argument and its covert subject overtly realised as the subject of the copula – is the sole topical (R8) argument of the copula verb.

Lumping various grammatical functions into single 'semantic roles' should be contrasted with splitting OBL into different 'semantic roles', on the basis of the preposition and the case it governs (see Tab. 3). For example, R2 – the approximation of Beneficiary and Recipient – is assigned not only to dative arguments, but also to arguments headed by DLA[gen] 'for', etc. Similarly, DO[gen] 'to', KU[dat] 'towards', NA[acc] 'on(to)', etc., are reasonable indicators of the Adlative role, approximated here by R4.

This algorithm ensures high uniqueness of 'semantic role' assignment. Out of the total number of 24 170 morphosyntactic schemata in the September 2013 version of Walenty, a valence dictionary for Polish (`http://zil.ipipan.waw.pl/Walenty`; Przepiórkowski *et al.* 2014), only 343 (or 1.42%) contained two or more arguments which would be mapped to the same 'semantic role'.[7] In almost half of them, namely 162, R4 would be duplicated; this is because the relevant schemata contain a number of prepositional arguments of the same type. This is also a problem for the underlying LFG grammar, as all such arguments need to be mapped to essentially the same grammatical function, OBL. As this would violate LFG's coherence condition, the grammar introduces also OBL2.[8] Exactly the same problem occurs with R6, which is duplicated 69 times. In case of the 48 duplicates of R8, valence schemata contain a broadly verbal argument (COMP or XCOMP) and one of the prepositional arguments listed in the R8 row of Tab. 3. Moreover, OBJ co-occurs with OBL-GEN or OBL-STR 35 times, resulting in the duplication of R1; the other 29 duplication cases are less systematic. For some of these 343 cases duplication cannot easily be avoided, but for others a more sophisticated 'semantic role' assignment procedure can be devised; e.g., when OBL-GEN occurs next to OBJ, it should probably be mapped to the broadly ablative R5 rather than the thematic R1.

Obviously, the procedure just described is an engineering heuristic, and instances of 'wrong' decisions may be found. For example, OBL arguments of type

[7] This should be contrasted with 2.47–4.36% of verb occurrences annotated with valence frames containing duplicated semantic roles in the experiments reported in Jaworski and Przepiórkowski 2014. As reported in that paper, on the same data the approach proposed here resulted in 1.73% of verb occurrences with valence frames containing duplicates.

[8] In fact, also OBL3 and OBL4. In the LFG valence dictionary, which was converted from the March 2014 version of Walenty, there are 19 787 schemata with OBL, 1843 (almost 10%) of them also mention OBL2, 45 of these 1843 include OBL3, and 2 of these 45 – also OBL4 (Agnieszka Patejuk, p.c.).

z[inst] ‘with’ have at least two meanings, apart from the perlative (R7): thematic (R1) and co-agentive (R0); in fact, the sentence *Zrób z nim porządek*, lit. ‘do with him order’, is ambiguous between the two and may mean either ‘Deal with him’ (R1) or ‘Clean up with him’ (R0).

On the positive side, while we do not have any quantitative data on the effects of this approach to ‘semantic roles’ on tasks such as textual entailment or question answering,[9] we note that it makes various inferences immediate which would not be straightforward if arguments were marked only with grammatical functions, e.g., inferences of the b. sentences from the corresponding a. sentences below:

(1) a. *Janek pobił Tomka.* ‘Janek beat Tomek up.’
b. *Tomek został pobity.* ‘Tomek was beaten up.’
(2) a. *Janek przesłał do Tomka książkę.* ‘Janek sent a book to Tomek.’
(lit. ‘Janek sent to Tomek (a/the) book.acc.’)
b. *Janek przekazał Tomkowi książkę.* ‘Janek transferred a book to Tomek.’
(lit. ‘Janek transferred Tomek.dat (a/the) book.acc.’)
(3) a. *Janek powiedział, że Tomek wygrał.* ‘Janek said that Tomek had won.’
b. *Janek mówił o Tomku.* ‘Janek was talking about Tomek.’

4 Conclusions

Given various problems with the practical applicability of standard repertoires of semantic roles reported in this paper, and the fact that creating resources such as VerbNet or FrameNet takes a lot of time, money and expertise, we proposed an ersatz solution consisting in assigning approximations of semantic roles calculated on the basis of syntactic (grammatical functions) and morphosyntactic (case, preposition form) features of arguments. The algorithm presented above makes it possible to assign such ‘semantic roles’ to arguments almost uniquely, and the resulting neo-Davidsonian representations facilitate textual entailments well beyond what would be possible if arguments were marked with grammatical functions only.

References

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In E. Hinrichs and K. Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT2002)*, Sozopol.

Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics. Blackwell, Malden, MA.

Bresnan, J. and Kanerva, J. M. (1989). Locative inversion in Chicheŵa: A case study of factorization in grammar. *Linguistic Inquiry*, **20**(1), 1–50.

[9] To the best of our knowledge, no testing data for such tasks are available for Polish and Polish has never been included in evaluation initiatives of this kind.

Copestake, A. and Flickinger, D. (2000). An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, Athens. ELRA.

Crouch, D. and King, T. H. (2005). Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes.*

Crouch, D. and King, T. H. (2006). Semantics via f-structure rewriting. In M. Butt and T. H. King, editors, *The Proceedings of the LFG'06 Conference*, Universität Konstanz, Germany. CSLI Publications.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2011). XLE documentation. `http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html`.

Dalrymple, M. (2001). *Lexical Functional Grammar.* Academic Press, San Diego, CA.

Dąbrowska, E. (1994). Dative and nominative experiencers: two folk theories of the mind. *Linguistics*, **32**, 1029–1054.

Dąbrowska, E. (1997). *Cognitive Semantics and the Polish Dative*, volume 9 of *Cognitive Linguistics Research.* Mouton de Gruyter, Berlin.

Fillmore, C. J. (1968). The case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88, New York. Holt, Rinehart and Winston.

Frank, A. (2004). Generalisations over corpus-induced frame assignment rules. In C. Fillmore, M. Pinkal, C. Baker, and K. Erk, editors, *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 31–38, Lisbon. ELRA.

Frank, A. and Erk, K. (2004). Towards an LFG syntax-semantics interface for frame semantics annotation. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing (CICLing 2004)*, volume 2945 of *Lecture Notes in Computer Science*, pages 1–12, Heidelberg. Springer-Verlag.

Frank, A. and Semecký, J. (2004). Corpus-based induction of an LFG syntax-semantics interface for frame semantic processing. In S. Hansen-Schirra, S. Oepen, and H. Uszkoreit, editors, *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora at COLING 2004*, Geneva.

Gruber, J. (1965). *Studies in Lexical Relations.* Ph.D. thesis, Massachusetts Institute of Technology.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar.* The MIT Press, Cambridge, MA.

Jakobson, R. O. (1971a). Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus. In *Selected Writings II*, pages 23–71. Mouton, The Hague.

Jakobson, R. O. (1971b). Morfologičeskie nabljudenija nad slavjanskim skloneniem. In *Selected Writings II*, pages 154–183. Mouton, The Hague.

Jaworski, W. and Przepiórkowski, A. (2014). Semantic roles in grammar engineering. In *Proceedings of the 3rd Join Conference on Lexical and Computational Semantics (*SEM 2014)*, Dublin, Ireland.

Kipper, K., Dang, H. T., Schuler, W., and Palmer, M. (2000). Building a class-based verb lexicon using TAGs. In *Proceedings of TAG+5 Fifth International Workshop on Tree Adjoining Grammars and Related Formalisms*.

Parsons, T. (1990). *Events in the Semantics of English: A Study in Subatomic Semantics*. The MIT Press, Cambridge, MA.

Patejuk, A. and Przepiórkowski, A. (2012). Towards an LFG parser for Polish: An exercise in parasitic grammar development. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3849–3852, Istanbul, Turkey. ELRA.

Pollard, C. and Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. Chicago University Press / CSLI Publications, Chicago, IL.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., and Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 2785–2792, Reykjavík, Iceland. ELRA.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.

Umemoto, H. (2006). Implementing a Japanese semantic parser based on glue approach. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 418–425, Huazhong Normal University, Wuhan, China. Tsinghua University Press.

Wierzbicka, A. (1980). *The Case for Surface Case*. Karoma, Ann Arbor, MI.

Wierzbicka, A. (1981). Case marking and human nature. *Australian Journal of Linguistics*, **1**, 43–80.

Wierzbicka, A. (1983). The semantics of case marking. *Studies in Language*, **7**, 247–275.

Wierzbicka, A. (1986). The meaning of a case: A study of the Polish dative. In R. D. Brecht and J. S. Levine, editors, *Case in Slavic*, pages 386–426. Slavica Publishers, Columbus, OH.

Draft of 4th June 2014– comments welcome! Don't cite without mentionting this date.