Ordering Slots of Semantically Related Schemata of Polish Verbs

Elżbieta Hajnicz
Institute of Computer Science, Polish Academy of Sciences
Elzbieta.Hajnicz@ipipan.waw.pl

Abstract
In this paper a method of ordering slots of verb schemata related by diathesis alternations is presented. Such schemata are grouped together and represent a single meaning of a verb. The schema dominating in a group is found and its slots are numbered w.r.t. their obliqueness hierarchy. These numbers are propagated to other schemata accordingly to alternations linking slots of schemata related by them.

1. Introduction
The primary task of our research is to create a semantic valence dictionary in an automatic way. To accomplish this goal, the valence dictionary of Polish verbs is supplemented with semantic information, provided by wordnet’s semantic categories (Hajnicz, 2009c) or synsets (Hajnicz, 2009a) of nouns. In our present work we focus on slots being nominal phrases and prepositional-nominal phrases, whose semantic heads are nouns. We discuss the case of 25 predefined semantic categories of nouns, which is simpler than the case of actual wordnet synsets.

In our previous works (Hajnicz, 2009b, 2010) we focused on the preparation of a syntactic-semantic valence dictionary, in which each slot of a syntactic schema is supplied with a list of semantic categories, forming a semantic frame. However, a genuine semantic dictionary is composed of semantic frames represented as a predicate-argument structure, in spite of their syntactic realisations. Each semantic argument is connected with its semantic role. On the other hand, syntactic slots of schemata are provided with information concerning which semantic role they realise.

A method of grouping schemata that participate in diathesis alternation was proposed in (Hajnicz, 2011a). However, the information that two schemata are semantically related is useless unless we know which slots carry the same semantic information (represent the same argument). Each group of schemata forms a single semantic dictionary entry, represented as a list of semantic arguments. Slots of every schema are linked to corresponding semantic arguments.

2. Related works
There exist several manually prepared semantic valence dictionaries. For English, the most famous are FrameNet (Baker et al., 2003; Fillmore et al., 2003; Ruppenhofer et al., 2010), VerbNet (Dang et al., 1998; Kipper et al., 2000; Kipper-Schuler, 2005; Kipper et al., 2006) and PropBank (Kingsbury and Palmer, 2002; Kingsbury et al., 2002; Palmer et al., 2005). In the case of Slavic languages, two extensive dictionaries for Czech, Vallex (Záboříčký and Lopatková, 2007) and VerbaLex (Hlaváčková and Horák, 2005, 2006) could be mentioned. Among these dictionaries, PropBank has abstract arguments ordered w.r.t. their obliqueness hierarchy (Keenan and Comrie, 1977, 1979; Croft, 2003), whereas VerbNet and VerbaLex have arguments interpreted by means of selectional preferences related to wordnet synsets.

3. Valence dictionary
A syntactic valence dictionary is a set of entries representing schemata for every verb considered. Each syntactic schema is a list of syntactic slots. The dictionary of 32 verbs chosen for the experiment was prepared on the basis of Świdziński’s dictionary (Świdziński, 1994). Verbs were chosen manually in a way to maximise the variability of their syntactic schemata (in particular, diathesis alternations) on one hand and polysyney within a single schema on the other. Their frequency was an important criterion for this choice as well.

The list of slots can include: adjectival phrases (AdjP), adverbial phrases (AdvP), infinitival phrases (InfP), nominal phrases (NP), prepositional-adjectival phrases (PrepAdjP), prepositional-nominal phrases (PrepNP) and clauses (SentP). A special slot sie hosts the reflexive marker. Some slots are parametrised. The only parameter of AdjP and NP is their case, the only parameter of InfP is its aspect. PrepAdjP and PrepNP have two parameters: the form of the preposition and the case of its AdjP or NP complement, respectively. SentP has one parameter, namely the complementizer introducing the clause. Below we list syntactic dictionary entries for the verb rozpocząć (to begin).

(1) rozpocząć advp np:nom
   rozpocząć np:acc np:inst np:nom
   rozpocząć np:acc np:nom
   rozpocząć np:acc np:nom preppnp:od:gen
   rozpoczęć np:acc np:nom preppnp:z:inst
   rozpocząć np:inst np:nom sie
   rozpocząć np:nom preppnp:dl:gen sie
   rozpocząć np:nom preppnp:od:gen sie
   rozpocząć np:nom sie

A syntactic-semantic valence dictionary was obtained by supplementing the syntactic valence dictionary with selectional preferences. Here, we consider the simple case of a fixed set of 25 semantic categories, which were assigned to nouns at the beginning of the preparation of
the Polish WordNet (Piancecki et al., 2009), which was modelled on the Princeton WordNet (Fellbaum, 1998) and wordnets constructed in the EuroWordNet project (Vossen, 1998).

4. Classification of alternations

In (Hajnicz, 2011b) we presented a very coarse, purely syntactic classification of potential alternations, describing only how the alternations relate slots in two schemata involved. The alternations can be divided into two types. First, there are alternations which preserve the number of slots in both schemata. This condition is satisfied by alternations referred to in (Hajnicz, 2011b) as simple alternation, exemplified by dative alternation, see (2), cross alternation exemplified by locative alternation, see (3), simple reflexive alternation, cf. (4), and cross-reflexive alternation, cf. (5). The sub-examples differ in the animacy of arguments.

(2) Chłopak posłał książkę koledże.
(A boy sent his friend a book.
Chłopak posłał książkę do kolegi.
(A boy sent a book to his friend)

(3) a. Trawa porosła wzgórze.
(Grass has grown over the hill.
Wzgórze porośla trawą.

b. Kwiaty pachną w ogrodzie.
(Flowers smell in the garden.
Ogród pachnie kwiatami/od kwiatów.
(The garden smells of flowers.

c. Rolnik ładuje wóz sianem.
(The farmer loads the wagon with hay.
Rolnik ładuje siano na wóz.
(The farmer loads hay onto the wagon.

(4) a. Chłopiec ogania muchy /śię od much.
(A boy drives away flies.
Chłopak kocha dziewczynę /śię w dziewczynie.
(A boy loves a girl.

(5) a. Hrabina urodziła syna.
(A countess gave birth to a son.
Syn urodził się hrabinie.
(A son was born to a countess.

b. Córka niepokoi matkę.
(A daughter worries (her) mother.
Matka niepokoi się córką /o córkę.
(Mother worries about (her) daughter.

c. Temat interesuje badacza.
(The subject is interesting to the researcher.
Badacz interesuje się tematem.
(The researcher is interested in the subject.

d. Publiczność wypełnił teatr.
(The audience filled the theatre.
Teatr wypełnił się publicznością.
(The theatre filled with the audience.

e. Wino napelnia kieliszkę.
(Wine fills glasses.
Kieliszkó napelniają się winem.
(Glasses fill with wine.

(6) Matka pożywiała naczynia.
(Mother washed the dishes.
Matka pożywiała.
(Mother washed)

(7) a. Jeździec pognął konia przez las.
(The rider rode a horse across a forest.
Konie pognął przez las.
(A horse rode across a forest.

b. Kelner napelnia kieliszkę winem.
(The waiter fills glasses with wine.
Wino napelnia kieliszek.
(Wine fills glasses.

(8) Żołnierz obronił towarzysza /śię przed atakiem.
(A soldier defended his comrade/himself from the attack.

(9) Chłopak spotkał dziewczynę /śię z dziewczyną.
(A boy met a girl.
Chłopak i dziewczynę spotkali się (ze sobą).
(A boy and a girl met (each other).

(10) a. Kelner złukł szklanki.
(A waiter broke glasses.
Szklanki złukły się.
(Glasses broke.

b. Kelner napelnia kieliszkę winem.
(The waiter fills glasses with wine.
Kieliszkó napelniają się winem.
(Glasses fill with wine.

c. Nadzorca zaharował niewolników (na śmierć).
(The overseer made slaves slog away (to death).
Niewolnicy zaharowali się (na śmierć).
(Staves slogged away (to death).

Semantically, reflexive deletion alternation preserves the number of slots, as się plays the role of the reflexive pronoun in the schema containing it carrying semantic information. Therefore, it is linked to a corresponding semantic argument in the semantic frame.

5. Rules of ordering slots

Our goal is to order verb slots in a group of schemata in a consistent way. Schemata in a group are semantically connected, which is accomplished by alternations relating them. The main idea is to find a dominant schema in

\footnote{These alternations cannot be differentiated at the level of schemata.}
a group, order its slots and subsequently propagate this
ordering to other schemata according to corresponding alternations.

Syntactic slots have the following priorities according to their obliqueness hierarchy, based on constructions as
Piotr przywiozl Annie kwiaty samochodem (Peter brought
Anna flowers by car., see also Ostler, 1979; Primus, 1999).

1. nominative np:nom,
2. accusative np:acc,
3. genitive np:gen,
4. dative np:dat,
5. instrumental np:inst,
6. prepositional phrases.

We have not employed any heuristics for ordering prepo-
sitional phrases not involved in alternations.

Each alternation contains information about which
slots should share the same semantic argument; slots not
involved in the alternation should agree both at the syntac-
tic and semantic level. Thus, in order to find a dominant
schema in the entire group we should establish the domi-
nant schema in every pair participating in the particular
alternation. For alternations “losing” slots, certainly a
schema which contains all slots is dominant. Therefore,
looking for the dominant schema, we only consider the
longest ones as candidates.

Depending on the thorough analysis of language mate-
rrial, we formulated the following heuristics of domination
of alternating schemata:

1. For simple and simple reflexive alternations, the
   schema in which the alternating slot has higher pri-
   ority dominates;
2. For cross alternation,
   (a) if one of alternating slots is PrepNP, then the
       schema containing it dominates (cf. (3) b., c.),
   (b) otherwise a schema in which the alternating slot
       has higher priority dominates (cf. (3) a.);
3. For reflexive cross alternation, domination depends
   on the animacy of the active or reflexive subject
   (a) if the reflexive object is dative, then the active
       schema dominates, (cf. (5) a.),
   (b) if the reflexive subject is animate, then the re-
       flexive schema dominates (cf. (5) b., c.),
   (c) if the reflexive subject is inanimate, then the
       active schema dominates (cf. (5) d.–f.),
   (d) if the subject is not involved in the alternation,\(^3\)
       then a schema in which the alternating slot has
       higher priority dominates.

This procedure enables us to find the dominating
schema in a group, order its slots and propagate this
information to other schemata.

Some of considered slots are adjuncts “typically re-
lated to some verbs and not to others” (Zabokrtský and

\(^3\)Such cases of reflexive cross alternation were not found in
the investigated set of verbs.

Lopatková, 2007). In PropBank such slots, usually prepo-
sitional, are put outside the obliqueness hierarchy and
labelled as ArgM (modifiers). We follow this conven-
tion for prepositional phrases not involved in any alter-
nation. Thus, slots having time, place, act or event as
their strongest selection preference are not ordered and
labelled with M instead.

6. Experiments

The process of ordering slots is concurrent with the
process of grouping schemata. The experiments were per-
formed using the manually prepared set of alternations
on the one hand (denoted as M) and the automatically
obtained set of alternations (Hajnizc, 2011b, denoted as
A). Alternations, in which the verb rozpocząć (to begin)
participates, are presented in (12). Semantically consis-
tent slots participating in the alternation are displayed
as np:nom, whereas semantically dropped ones are dis-
played as np:nom. All schemata of the verb rozpocząć
are semantically related and form a single group (11).

\begin{align*}
(11)\quad \text{rozpocząć} & \quad 1 \quad \text{A1, A2, A3} \\
& \quad \text{np:acc:A2} \quad \text{np:inst:A3} \quad \text{np:nom:A1} \\
& \quad \text{np:acc:A2} \quad \text{np:nom:A1} \\
& \quad \text{np:acc:A2} \quad \text{np:nom:A1} \quad \text{prepnp:od:gen:A3} \\
& \quad \text{np:inst:A3} \quad \text{np:nom:A2} \quad \text{sie} \\
& \quad \text{np:nom:A2} \quad \text{prepnp:dla:gen:A1} \quad \text{sie} \\
& \quad \text{np:nom:A2} \quad \text{prepnp:od:gen:A3} \quad \text{sie} \\
& \quad \text{np:nom:A2} \quad \text{sie}
\end{align*}

Experiments were conducted in three ways. First, all
slots were ordered. Secondly, two factors \(F\) of preference
strength were considered: 1 and 1.5, meaning that “ad-
junctive” selection preference is \(F\) times greater than all
other preferences in sum. As a baseline (denoted as B),
ordering of slots was performed independently for each
schema.

Evaluation was performed using a semantic dictionary
prepared manually especially for our experiments. Its ex-
emplary entry is presented in (11). The simplest way of
evaluation would be to check for the exact match of la-
beles. However, the results of grouping schemata could
influence the label assignment. In order to evaluate label
assignment exclusively, we decided to accept labelling as
correct if ordering is preserved, i.e., all slots preceding
and succeeding the validated one in the automatically ob-
tained dictionary precede and succeed it in the manually
prepared dictionary.

In spite of the above mentioned differences, each slot
could be correctly labelled or not, there are no other pos-
sibilities. Thus, the only appropriate measure is correct-
ness. We calculate it for particular slots and for the whole
schema, meaning that all their slots are properly ordered.
Correctness was calculated in two ways. First, in a per-
missive way (denoted as \(\Omega\)), in which modifier M agrees
with any actual argument A1. Secondly, in a rigorous way
(denoted as \(\cap\)), in which all modifiers M must match.

The results of evaluation are presented in Table 1. Symbol \(\infty\) represents the infinitive factor of the “ad-
junctive” preference strength, i.e., ordering all slots. In this
case, only the permissive calculation of correctness is ap-
reflexive shift
\textsf{np:acc np:inst np:nom} \quad \textsf{np:inst np:nom sie}
\textsf{np:acc np:nom} \quad \textsf{np:nom sie}
\textsf{np:acc np:nom prepnp:od:gen} \quad \textsf{np:nom prepnp:od:gen sie}

simple
\textsf{np:acc np:inst np:nom} \quad \textsf{np:acc np:nom prepnp:od:gen}
\textsf{np:inst np:nom sie} \quad \textsf{np:nom prepnp:od:gen sie}
\textsf{np:nom prepnp:dia:gen sie} \quad \textsf{np:nom sie}
\textsf{np:nom prepnp:dia:gen sie} \quad \textsf{np:nom sie}

Table 1: Evaluation of labelling slots in a semantic dictionary

<table>
<thead>
<tr>
<th>data</th>
<th>slots</th>
<th>schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M^1)</td>
<td>99.89</td>
<td>99.89</td>
</tr>
<tr>
<td>(M^2)</td>
<td>99.42</td>
<td>99.29</td>
</tr>
<tr>
<td>(A^1)</td>
<td>97.44</td>
<td>97.69</td>
</tr>
<tr>
<td>(A^\cap)</td>
<td>96.63</td>
<td>97.09</td>
</tr>
<tr>
<td>(B^1)</td>
<td>96.93</td>
<td>98.69</td>
</tr>
<tr>
<td>(B^\cap)</td>
<td>96.12</td>
<td>80.08</td>
</tr>
</tbody>
</table>

The results of evaluation are impressive, being good even for the baseline. Observe that factor 1 considers too many slots as modifiers, which results in the worst restrictive evaluation. Thus, factor 1.5 should be considered optimal. Permissive evaluation depends on the choice of modifiers rather weakly, only for the automatically detected alternations. This means that modifiers only mask errors in the numbering of slots, which manifests itself in a difference between permissive and restrictive evaluation, greater than for manually prepared alternations.

Results for automatically detected alternations are 2-3 percentage points worse than results for manually prepared alternations. This means that the method is sensitive to data errors. Still, they are better than the baseline, especially if modifiers are detected. The reason is that the baseline method considers all prepositional phrases as potential modifiers, having no information about their participation.

7. Conclusions and future work

In this paper, a method of consistent semantic labelling of slots in a group of semantically related verb schemata was proposed. It is based on several heuristics concerning the obliqueness hierarchy of slots in a schema and domination of schemata related by a particular diathesis alternation.

The experiments performed on a set of verbs larger than the present one of 32 verbs would give more reliable results. However, the chosen verbs have proportionally large sets of related schemata, hence the task of ordering their slots is probably harder than in average.

The results are so good that the method could be applied fully automatically, especially for manually prepared or corrected sets of alternations. This by no means concerns other steps of creating a semantic valence dictionary. Thus, further efforts should be devoted to the improvement of methods of alternation detection and grouping schemata. Nevertheless, the whole process could be performed only semi-automatically. The automatic detection of selectional preferences, alternations and automatic grouping of schemata will be a valuable support for lexicographers creating such a dictionary.

(13) rozpocząć
A1  person=0.8, group=0.2
A2  act=0.45, event=0.3, time=0.1
A3  event=0.4, act=0.3, artifact=0.2

Semantic dictionary entries, like (11), representing particular verb meanings, are assigned to sets of semantically related schemata of verbs, with the uniform numbering of slots. Such an entry, together with selectional preferences of arguments, as in (13), could be a valuable source of information about their semantic roles. This is an interesting subject for further research as well.

Acknowledgements This research is supported by the POIG.01.01.02-14-013/09 project which is co-financed by the European Union under the European Regional Development Fund.

References


