Argument co-occurrence matrix as a description of verb valence

Łukasz Dębowski, Marcin Woliński

Instytut Podstaw Informatyki PAN
J.K. Ordona 21, 01-237 Warszawa, Poland
{ldebowsk,wolinski}@ipipan.waw.pl

Abstract

A new description of verb valence is proposed. Rather than by full valence frames, verb subcategorization is described by the list of arguments and the matrix of five distinct pairwise argument interactions. This approach is argued to be computationally robust and sufficient although it has been initially motivated only by the need of a reliable test set for machine verb valence learning. Speaking more abstractly, we propose an approach to decomposing empirical subsets of powersets that reduces data sparseness problems and an approach to discretize (summarize) a large number of contingency tables. Both techniques are novel and might be useful not only in linguistics.

1. Introduction

In this paper,[1] we propose a new representation of verb valence frames, namely a co-occurrence matrix for verb arguments. This kind of representation seems:

1. detailed enough to represent verb valence frames concisely, extensively, and fairly sufficiently,
2. robust enough to be acquired with high agreement rate by linguists compiling verb valence dictionaries,
3. easy to use in sentence parsing e.g. via metamorphosis grammars (c.f. e.g. Colmerauer, 1978; Świdziński, 1992; Woliński, 2004),
4. simple enough for automatic verb valence acquisition via machine learning (c.f. e.g. Fast and Przepiórkowski, 2005; Przepiórkowski, 2006).

In the following, we will briefly describe our approach. Advantages and drawbacks of co-occurrence matrices will be compared to the features of presently used schemes for verb valence dictionaries. The discussion is developed from a computational perspective and our concepts are illustrated on Polish language examples.

The concept of co-occurrence matrix seems to be proposed for the first time and may be useful not only in computational linguistics. Therefore this paper details the motivation and formal construction of co-occurrence matrices. For lack of space, we have not addressed some questions that concern learning of such matrices.

2. Inadequacies of valence dictionaries

It is commonly recognized that certain types of dependent phrases (arguments) can be combined with a given verb, whereas other types of phrases may not. E.g.:

(1) Jarek jest aktorem. (=Jarek is an actor.)

(2) *Jarek jest ciemno. (≈*Jarek is darkly.)

In the example, Polish verb *być* (to be) requires two arguments: (i) the subject in nominative and (ii) the attribute in instrumental case (if it is a noun). We can represent this information by verb argument list $\mathbf{L}(v)$ like (14)–(15) in Table 4. Such approach is followed by Mędak (2005).

Trying to describe the linguistic reality more precisely, a researcher soon realizes that certain arguments exclude mutually or imply one another. Compare (1), (3), and (4):

(3) Jarek jest doskonały. (=Jarek is perfect.)

(4) *Jarek jest aktorem doskonały.

It is impossible to use *być* with np(inst) and adjp(nom) phrases simultaneously. Some reaction to this phenomenon is compiling verb valence dictionaries as sets of valence frames $\mathbf{F}(v) \subset 2^{\mathbf{L}(v)}$, such as (16)–(17) in Table 4.[2] This approach was carried out by Polański (1980–1992), Świdziński (1994), and Bańko (2000). Seemingly any verb valence could be thus described given enough patience, effort, and thoroughness.

The expressibility of scheme (16)–(17), however, is a source of numerous mistakes, omissions, and redundancies in the manually compiled dictionaries. For instance, researchers that strive for describing argument dropping (ellipsis) try to list all incomplete valence frames in the $\mathbf{F}(v)$ set. Nevertheless, for N different types of arguments that can be combined with a verb, there are 2^N distinct subsets of these arguments. Potentially, each of the subsets must be considered by the lexicographer and determined to be or not to be a valid valence frame.

It is not surprising that hand-crafted valence dictionaries, such as Polański, Świdziński, and Bańko, differ substantially with respect to $\mathbf{F}(v)$ (cf. Table 1 here, as well as Przepiórkowski and Fast (2005)). The disagreement is far less when the first valence function

$$\mathbf{L}(v) := \left\{ i \mid \exists_{f \in \mathbf{F}(v)}\, i \in f \right\} \tag{5}$$

is considered. Especially, the precision of the dictionaries is significantly greater for $\mathbf{L}(v)$ than for $\mathbf{F}(v)$.[3] We have

[1] The work was supported by the Polish State Research Project, 3 T11C 003 28, *Automatyczna ekstrakcja wiedzy lingwistycznej z dużego korpusu języka polskiego*.

[2] We assume for simplicity that no argument type can be repeated in a valence frame of a verb. This restriction can be overcome by assigning unique identifiers to the repeated arguments.

[3] This empirical fact cannot be deduced from mere (5). For too little space, we skip over the constructed counterexample.

F	Bań.	Pol.	Świ.	MV
Bań.	554			
Pol.	238	481		
Świ.	266	247	425	
MV	312	293	321	367
recall	0.85	0.8	0.87	
precision	0.56	0.61	0.76	

L	Bań.	Pol.	Świ.	MV
Bań.	513			
Pol.	329	467		
Świ.	348	341	470	
MV	383	376	395	430
recall	0.89	0.87	0.92	
precision	0.75	0.81	0.84	

Table 1: A comparison of dictionaries by Bańko (2000), Polański (1980–1992), and Świdziński (1994), marked Bań., Pol., and Świ. For each dictionary, we computed sets $\mathbf{F} := \{(v,f) \mid v \in V, f \in \mathbf{F}(v)\}$ and $\mathbf{L} := \{(v,i) \mid v \in V, i \in \mathbf{L}(v)\}$, where V was a sample of 95 verbs. Besides we compiled the majority voting (MV) versions of $\mathbf{L}$ and $\mathbf{F}$. The table presents cardinalities of the intersections of $\mathbf{L}$'s and $\mathbf{F}$'s for each pair of dictionaries. Recall and precision are given against the MV version. We ignored some abstract arguments used by Polański (i.e., those not defined formally).

checked that the agreement of the dictionaries varies with respect to specific arguments. Overt subjects, transitivity, and directional arguments are marked coherently but there is less agreement about action beneficiaries, instruments, obligatory adverbs, or even sentential phrases. The statistics are omitted for lack of space.

Unfortunately, one cannot expect that the larger a dictionary the larger the number of reliable entries it provides. Both Polański and Bańko present more valence frames than Świdziński. Nevertheless, the absolute number of frames shared by Polański and Bańko is smaller than the number of frames shared by any of them and Świdziński. The same holds if single arguments are considered instead of frames (cf. Table 1).

Since human-made dictionaries appear superficially inconsistent, it seems hard to provide a gold standard test set for machine valence dictionary learning (cf. Przepiórkowski and Fast, 2005). Acquiring schemes like (16)–(17) from texts seems hopeless as well since too many elements of $\mathbf{F}(v)$ do not occur by chance in the empirical data for many a verb v. In order to find a reliable gold standard and to design a robust learning procedure, one has to reduce the number of learned parameters but reducing the verb valence to $\mathbf{L}(v)$ is too much.

3. Argument co-occurrence matrix

Halford et al. (1998) suggested that human working memory limits may be plausibly defined in terms of the maximal arity of a relation that can be processed as a single entity. They claimed that the human mind preferably approximates more complex relations by some logical formulae involving relations of low arity. The set of preferred relations does not include even all relations of fixed low arity. For instance, analogies (Lepage, 2001, Section 2.2) are the most preferred quaternary relations.

Having the above in mind, we are skeptical about the idea that verb valence $\mathbf{F}(v)$ can be an arbitrary subset of $2^{\mathbf{L}}(v)$. Instead, we propose decomposing $\mathbf{F}(v)$ through its projections onto simpler argument relations. We can start with binary projections. For a fixed verb v, let $\langle i \rangle := \{f \in \mathbf{F}(v) \mid i \in f\}$ be the set of verb frames which contain argument type i. The intuition that some arguments of verb v exclude or imply one another can be formally represented by co-occurrence matrix $\mathbf{M}(v) : \mathbf{L}(v) \times \mathbf{L}(v) \to \{\leftarrow, \to, \leftrightarrow, \times, \perp\}$ constructed from sets $\langle i \rangle$. We will put

$$\mathbf{M}(v)_{ij} := \mathrm{R} \iff i\,\mathrm{R}\,j \tag{6}$$

where the implicitly verb-dependent relations are

$$i \times j \iff \langle i \rangle \cap \langle j \rangle = \emptyset, \qquad (i \text{ excludes } j) \tag{7}$$

$$i \leftrightarrow j \iff \langle i \rangle = \langle j \rangle, \qquad (i \text{ and } j \text{ co-occur}) \tag{8}$$

$$i \to j \iff [\langle i \rangle \subset \langle j \rangle \wedge \langle i \rangle \neq \langle j \rangle], \qquad (i \text{ implies } j) \tag{9}$$

$$i \leftarrow j \iff [\langle i \rangle \supset \langle j \rangle \wedge \langle i \rangle \neq \langle j \rangle], \qquad (j \text{ implies } i) \tag{10}$$

$$i \perp j \iff [\langle i \rangle \setminus \langle j \rangle, \langle i \rangle \cap \langle j \rangle, \langle j \rangle \setminus \langle i \rangle \neq \emptyset]. \tag{11}$$

Symbol $\perp$ that denotes "formal" independence was chosen intentionally as resembling symbol $\perp\!\!\!\perp$, which is usually applied to mean probabilistic independence.

In the examples of $\mathbf{M}(v)$, given in Tables 2 and 4(18), we write $\leftarrow$, $\upharpoonright$, and ⇷ instead of $\leftarrow$, $\to$, and $\leftrightarrow$ for pictorial clarity. The matrix is symmetric in a generalized sense if we agree that $\leftarrow$ and $\upharpoonright$ are reflections of each other.

For the previously considered hand-crafted dictionaries, the agreement about $\mathbf{M}(v)$ is quite high (cf. Table 3). If argument types i and j are listed as valid arguments for verb v in two dictionaries then there is 85%–90% chance that relation $\mathbf{M}(v)_{ij}$ is the same in both dictionaries. The greatest source of confusion is seemingly failing to recognize weak independence $\perp$. Exclusion $\times$ and mutual implication $\leftrightarrow$ are almost never confused with each other. The analogical statement applies to pairs $\{\leftarrow, \to\}$ and $\{\leftrightarrow, \perp\}$.

4. Parsing with the co-occurrence matrix

Define the set of obligatory verb arguments

$$\mathbf{E}(v) := \left\{ i \in \mathbf{L}(v) \mid \forall_{f \in \mathbf{F}(v)}\, \psi(f,i) \right\}.$$

Having just matrix $\mathbf{M}(v)$ and sets $\mathbf{L}(v)$ and $\mathbf{E}(v)$, we may try to reconstruct $\mathbf{F}(v)$ in some approximation. First, define predicate ϕ as

$$\phi(f,\mu,i,j) := \begin{cases} \neg(\psi(f,i) \wedge \psi(f,j)), & \mu_{ij} = \times, \\ \psi(f,i) \iff \psi(f,j), & \mu_{ij} = \leftrightarrow, \\ \psi(f,i) \implies \psi(f,j), & \mu_{ij} = \to, \\ \psi(f,i) \impliedby \psi(f,j), & \mu_{ij} = \leftarrow, \\ \text{true}, & \mu_{ij} = \perp \end{cases}$$

for $\psi(f,i) := (i \in f)$. Following this we can propose an intuitive reconstruction of $\mathbf{F}(v)$, namely

$$\bar{\mathbf{F}}(v) := \left\{ f \in 2^{\mathbf{L}(v)} \,\middle|\, \begin{array}{l} \forall_{i \in \mathbf{E}(v)}\, \psi(f,i), \\ \forall_{i,j \in \mathbf{L}(v)}\, \phi(f, \mathbf{M}(v), i, j) \end{array} \right\}.$$

	np(nom)	np(acc)	advp	np(dat)	np(inst)	pp(w,loc)	pp(do,gen)	pp(na,acc)	pp(z,gen)	sentp(że)
np(nom)	$\leftarrow\!\!\!\!\uparrow$:95/95	←:47/71	←:49/52	←:43/47	←:45/46	←:32/35	←:30/32	←:27/28	←:23/23	←:21/23
np(acc)	↑:47/71	$\leftarrow\!\!\!\!\uparrow$:71/71	←:20/35	←:17/34	←:23/40	←:14/23	←:10/21	←:14/22	←:11/17	×:13/17
advp	↑:49/52	↑:20/35	$\leftarrow\!\!\!\!\uparrow$:52/52	×:21/29	×:23/27	×:23/25	×:24/26	×:22/23	×:15/16	×:12/14
np(dat)	↑:43/47	↑:17/34	×:21/29	$\leftarrow\!\!\!\!\uparrow$:47/47	×:14/22	×:12/17	×:16/20	×:17/20	×:10/11	⊥:7/11
np(inst)	↑:45/46	↑:23/40	×:23/27	×:14/22	$\leftarrow\!\!\!\!\uparrow$:46/46	×:17/20	×:16/18	×:12/14	×:15/15	×:9/9
pp(w,loc)	↑:32/35	↑:14/23	×:23/25	×:12/17	×:17/20	$\leftarrow\!\!\!\!\uparrow$:35/35	×:11/11	×:9/10	×:11/11	×:7/7
pp(do,gen)	↑:30/32	↑:10/21	×:24/26	×:16/20	×:16/18	×:11/11	$\leftarrow\!\!\!\!\uparrow$:32/32	×:19/19	×:11/12	×:4/5
pp(na,acc)	↑:27/28	↑:14/22	×:22/23	×:17/20	×:12/14	×:9/10	×:19/19	$\leftarrow\!\!\!\!\uparrow$:28/28	×:10/10	×:5/7
pp(z,gen)	↑:23/23	↑:11/17	×:15/16	×:10/11	×:15/15	×:11/11	×:11/12	×:10/10	$\leftarrow\!\!\!\!\uparrow$:23/23	⊥:1/1
sentp(że)	↑:21/23	×:13/17	×:12/14	⊥:7/11	×:9/9	×:7/7	×:4/5	×:5/7	⊥:1/1	$\leftarrow\!\!\!\!\uparrow$:23/23

Table 2: The most frequent cell values of co-occurrence matrices for the set-theoretic sum of Bań., Pol., and Świ. (constrained to the sample of 95 verbs). For each pair of arguments we give x:y/z with x — the most frequent relation, y — the number of verbs for which x is satisfied, and z — the total number of verbs that allow both arguments. The leading columns and rows list top 10 most frequent arguments and their unigram frequencies are given respectively in decreasing order on the diagonal.

This reconstruction is the maximal set of frames that contain all the required arguments and induce $\mathbf{M}(v)$ as their co-occurrence matrix.

Notice that our redefinition of verb valence can be readily used for sentence parsing. Typically, the parser checks whether a hypothetical frame f of a parsed sentence belongs to $\mathbf{F}(v)$ (cf. Woliński, 2004). We propose to check only whether $f \in \bar{\mathbf{F}}(v)$. This results in parser accepting more sentences since $\bar{\mathbf{F}} \supset \mathbf{F}$.[4]

5. Extended arguments and their matrix

Now let us discuss a more acute deficiency of defining verb valence frames with binary relations. The deficiency has to do with an interaction between the argument exclusion and argument implication. The simplest example can be provided by verb *być*. According to (16), the verb requires the presence of either np(nom) or advp. Nevertheless these arguments cannot co-occur so we have (19).

Interaction between exclusion and implication causes troubles when the mutually exclusive arguments are implied rather than implying. Compare (20)–(21). Verb *pożyczyć* (to lend/borrow), although polysemous, presents no problems. Relations

$$\begin{array}{ll} \text{np(nom)} \leftrightarrow \text{np(acc)}, & \text{np(dat)} \times \text{pp(od,gen)}, \\ \text{np(dat)} \rightarrow \text{np(acc)}, & \text{pp(od,gen)} \rightarrow \text{np(acc)} \end{array}$$

describe the valence exactly, that is, $\bar{\mathbf{F}}(v) = \mathbf{F}(v)$ for $v =$ *pożyczyć*. The problem arises for $v =$ *powiedzieć* (to say). Arguments np(acc) and sentp(że) exclude each other but an instance of np(dat) requires either of them.[5]

Thus, $\mathbf{M}(v)_{\text{np(dat},j)} = \perp$ for $j = \text{np(acc)}, \text{sentp(że)}$ and $\bar{\mathbf{F}}(v) \neq \mathbf{F}(v)$ consequently.

We can, however, propose another improvement. Define a set of extended arguments

$$\mathbf{L}^*(v) := \left\{ k \in 2^{\mathbf{L}(v)} \;\middle|\; \begin{array}{l} k \neq \emptyset, \\ \forall_{i,j \in k}\, i \sim j \end{array} \right\}, \tag{12}$$

where $i \sim j \iff (i \times j \vee i = j)$. The set contains all arguments of the verb (as singletons) and all subsets of mutually exclusive arguments.

Rather than using matrix $\mathbf{M}(v)$, we will construct an $\mathbf{L}^*(v) \times \mathbf{L}^*(v)$ matrix $\mathbf{M}^*(v)$ with cells

$$\mathbf{M}^*(v)_{kl} := \mathrm{R} \iff k \,\mathrm{R}\, l,$$

where the relations are defined by (7)–(11) with $\langle k \rangle := \{f \in \mathbf{F}(v) \mid k \cap f \neq \emptyset\}$. Condition $\psi^*(f,k) := (k \cap f \neq \emptyset)$ plays role of the previous $\psi(f,i)$. Therefore, the analog of $\mathbf{E}(v)$ is

$$\mathbf{E}^*(v) := \left\{ k \in \mathbf{L}^*(v) \mid \forall_{f \in \mathbf{F}(v)}\, \psi^*(f,k) \right\}.$$

Finally, define reconstruction

$$\bar{\mathbf{F}}^*(v) := \left\{ f \in 2^{\mathbf{L}(v)} \;\middle|\; \begin{array}{l} \forall_{k \in \mathbf{E}^*(v)}\, \psi^*(f,k), \\ \forall_{k,l \in \mathbf{L}^*(v)}\, \phi^*(f, \mathbf{M}^*(v), k, l) \end{array} \right\}, \tag{13}$$

where predicate ϕ^* is defined as ϕ with ψ^* replaced for ψ.

The new objects are related to the previous through $\mathbf{F} \subset \bar{\mathbf{F}}^* \subset \bar{\mathbf{F}}$ and $\mathbf{M}^*(v)_{\{i\},\{j\}} = \mathbf{M}(v)_{ij}$. Of course, $\{i\} \in \mathbf{E}^*(v)$ is equivalent to $i \in \mathbf{E}(v)$.

Some examples of sets $\mathbf{L}^*(v)$ are (22)–(23). Although the sets of extended arguments can be big, we achieve the correct sets of required alternative arguments (24). We also obtain $\{\text{np(dat)}\} \rightarrow \{\text{np(acc)}, \text{sentp(że)}\}$ for *powiedzieć*, which accomplishes our initial goal.

6. Efficient check of the extension

The set of extended arguments can be computed relatively fast given $\mathbf{L}(v)$ and $\mathbf{M}(v)$ if dynamic programming is applied. The naive solution suggested by formula (12)

[4] $\bar{\mathbf{F}}(v)$ can be modified easily to make the parser accept also elliptic utterances. Just redefine $\psi(f,i) :=$ true and

$$\phi(f,\mu,i,j) := \begin{cases} \neg(\psi(f,i) \wedge \psi(f,j)), & \mu_{ij} = \times, \\ \text{true}, & \text{else.} \end{cases}$$

[5] Argument exclusion is usually connected to a kind of semantic equivalence of the arguments. Namely, the mutually exclusive arguments convey the same type of information and they should be coordinated rather than concatenated.

		Bań.						
		×	←	→	↔	⊥	N/A	Σ
Pol.	×	328	3	3	–	28	788	1150
	←	8	223	–	7	18	207	463
	→	8	–	223	7	18	207	463
	↔	2	23	23	383	2	144	577
	⊥	24	11	11	–	30	126	202
	N/A	1260	257	257	184	102		
	Σ	1630	517	517	581	198		

	Σ	same	different	agreement rate
in both sources	1383	1187	196	0.86
only in Pol.	1472			
only in Bań.	2060			

		Bań.						
		×	←	→	↔	⊥	N/A	Σ
Świ.	×	364	2	2	–	30	724	1122
	←	4	253	1	2	18	176	454
	→	4	1	253	2	18	176	454
	↔	–	25	25	410	2	124	586
	⊥	16	6	6	–	28	38	94
	N/A	1242	230	230	167	102		
	Σ	1630	517	517	581	198		

	Σ	same	different	agreement rate
in both sources	1472	1308	164	0.89
only in Świ.	1238			
only in Bań.	1971			

		Pol.						
		×	←	→	↔	⊥	N/A	Σ
Świ.	×	340	8	8	–	52	714	1122
	←	2	241	1	9	14	187	454
	→	2	1	241	9	14	187	454
	↔	–	12	12	427	–	135	586
	⊥	6	7	7	2	34	38	94
	N/A	800	194	194	130	88		
	Σ	1150	463	463	577	202		

	Σ	same	different	agreement rate
in both sources	1449	1283	166	0.89
only in Świ.	1261			
only in Pol.	1406			

Table 3: Three larger subtables present the numbers of triplets (v, i, j), where $v \in V$, $i, j \in \mathbf{L}(v)$, and $\mathbf{M}(v)_{ij}$ is equal to the specified relations according to the chosen pair of dictionaries. If triplet (v, i, j) appears only in one dictionary, it is counted as N/A for the other dictionary. The total numbers of triplets (v, i, j) are given in three smaller subtables. Triplets for which both dictionaries give the same value of $\mathbf{M}(v)_{ij}$ are counted as "same". The other ones are counted as "different".

is to search through all elements of power set $2^{\mathbf{L}(v)}$ and to check for each independently whether it is an element of $\mathbf{L}^*(v)$. We can, however, act smarter. Enumerate the elements of $\mathbf{L}(v) = \{i_1, i_2, ..., i_N\}$ and compute iteratively

$$A_0 = \{\emptyset\}\,,$$
$$A_n = A_{n-1} \cup \{\{i_n\} \cup k \mid k \in A_{n-1}, \forall_{i \in k}\, i \sim i_n\}\,,$$

where $n = 1, 2, ..., N$. In fact, $A_N = \mathbf{L}^*(v)$ and $A_n = \{k \in \mathbf{L}^*(v) \mid \forall_{p>n}\, i_p \notin k\}$. Analogous iterations may be formulated also for reconstructions $\bar{\mathbf{F}}(v)$ and $\bar{\mathbf{F}}^*(v)$.

Because the number of extended arguments is surprisingly large for certain verbs (such as (22)), it is natural to ask whether all these arguments are necessary. One can seek for a smaller set $\mathbf{L}^R(v) \subset \mathbf{L}^*(v)$ that can be substituted for $\mathbf{L}^*(v)$ in the left-hand side of (13).

Notice that if an extended argument belongs to $\mathbf{L}^*(v)$ then the singletons containing each of its elements belong to $\mathbf{L}^*(v)$, as well. Thus we can ask a simple question:

> Which values of $\mathbf{M}^*(v)_{kl}$ can be predicted given $\mathbf{M}^*(v)_{\{i\}l}$ for all $i \in k$?

The set of predictable relations is $\mathbf{P} = \{\times, \rightarrow, \perp\}$ since

$$\mathbf{M}^*(v)_{kl} = \times \iff \forall_{i \in k}\ \{i\} \times l,$$
$$\mathbf{M}^*(v)_{kl} = \rightarrow \iff \forall_{i \in k}\ \{i\} \rightarrow l,$$
$$\mathbf{M}^*(v)_{kl} = \perp \iff \begin{cases} \mathbf{M}^*(v)_{kl} \notin \{\times, \rightarrow\}\,, \\ \forall_{i \in k}\, \mathbf{M}^*(v)_{\{i\}l} \in \mathbf{P}. \end{cases}$$

On the other hand, given $\mathbf{M}^*(v)_{\{i\}l}$, we cannot predict the value of $\mathbf{M}^*(v)_{kl}$ if $k \leftarrow l$ or $k \leftrightarrow l$ for $k \not\supset l$.

Hence define a set of extended non-singleton arguments exhibiting the predictable or trivial behavior as

$$\mathbf{L}^P(v) = \left\{ k \in \mathbf{L}^*(v) \,\middle|\, \begin{array}{l} \operatorname{card} k > 1,\ k \notin \mathbf{E}^*(v), \\ \forall_{l \in \mathbf{L}^*(v)}\, k \supset l \vee \mathbf{M}^*(v)_{kl} \in \mathbf{P} \end{array} \right\}.$$

Observe that, as we wanted, $\mathbf{L}^*(v)$ in equation (13) can be replaced with the reduced set of extended arguments

$$\mathbf{L}^R(v) = \mathbf{L}^*(v) \setminus \mathbf{L}^P(v).$$

An example of set $\mathbf{L}^R(v)$ is (25), whereas the original set of extended arguments (22) is more than twice larger.

7. Machine learning issues

By the time of submitting this article, we have not elaborated the detailed scheme of machine learning for cooccurence matrices. Nevertheless, designing robust learning procedures for $\mathbf{M}^*(v)$ and $\mathbf{L}^*(v)$ seems much easier than for full valence frames $\mathbf{F}(v)$.

For instance, notice that the value of $\mathbf{M}^*(v)_{kl}$ can be obtained by discretizing a single contingency table

	$\neg\psi^*(\cdot, k)$	$\psi^*(\cdot, k)$
$\neg\psi^*(\cdot, l)$	$N - N_k - N_l + N_{kl}$	$N_k - N_{kl}$
$\psi^*(\cdot, l)$	$N_l - N_{kl}$	N_{kl}

where $N = \operatorname{card} \mathbf{F}(v)$, $N_k = \operatorname{card} \langle k \rangle$, $N_l = \operatorname{card} \langle l \rangle$, and $N_{kl} = \operatorname{card}(\langle k \rangle \cap \langle l \rangle)$. Relations ←, →, ↔ and × correspond to structural zeroes in certain cells of the table (Mohri and Roark, 2005).

Secondly, accordingly to Table 2, relations $\mathbf{M}^*(v)_{kl}$, k and l fixed, are highly similar for different verbs. We may suppose that also the non-singleton elements of $\mathbf{L}^R(v)$ (i.e. extended arguments) re-instantiate with many verbs.

Having the above two factors in mind, it should not be hard to devise reliable learning procedures based on a mix of bootstrap Monte Carlo tests (Durka, 2003) and Bayesian inference. Details of this issue, with more references, will be provided in a longer technical report.

(14) $\mathbf{L}(\textit{być}) = \{\text{np(nom)}, \text{np(inst)}, \text{adjp(nom)}, \text{infp}, \text{advp}\}$ (=to be).

(15) $\mathbf{L}(\textit{dać}) = \{\text{np(nom)}, \text{np(acc)}, \text{np(dat)}, \text{infp}\}$ (=to give/allow).

(16) $\mathbf{F}(\textit{być}) = \{\{\text{np(nom)}\}, \{\text{np(nom)}, \text{np(inst)}\}, \{\text{np(nom)}, \text{adjp(nom)}\}, \{\text{advp}\}, \{\text{infp}, \text{advp}\}\}$.

(17) $\mathbf{F}(\textit{dać}) = \{\{\text{np(nom)}, \text{np(acc)}, \text{np(dat)}\}, \{\text{np(nom)}, \text{np(dat)}, \text{infp}\}\}$.

(18)

$\mathbf{M}(\textit{być})$	np(nom)	np(inst)	adjp(nom)	infp	advp
np(nom)	$\leftarrow\!\!\!\!\!\uparrow$	$\leftarrow$	$\leftarrow$	$\times$	$\times$
np(inst)	$\uparrow$	$\leftarrow\!\!\!\!\!\uparrow$	$\times$	$\times$	$\times$
adjp(nom)	$\uparrow$	$\times$	$\leftarrow\!\!\!\!\!\uparrow$	$\times$	$\times$
infp	$\times$	$\times$	$\times$	$\leftarrow\!\!\!\!\!\uparrow$	$\uparrow$
advp	$\times$	$\times$	$\times$	$\leftarrow$	$\leftarrow\!\!\!\!\!\uparrow$

$\mathbf{M}(\textit{dać})$	np(nom)	np(acc)	np(dat)	infp
np(nom)	$\leftarrow\!\!\!\!\!\uparrow$	$\leftarrow$	$\leftarrow$	$\leftarrow$
np(acc)	$\uparrow$	$\leftarrow\!\!\!\!\!\uparrow$	$\uparrow$	$\times$
np(dat)	$\uparrow$	$\leftarrow$	$\leftarrow\!\!\!\!\!\uparrow$	$\leftarrow$
infp	$\uparrow$	$\times$	$\uparrow$	$\leftarrow\!\!\!\!\!\uparrow$

(19) $\mathbf{E}(\textit{być}) = \emptyset, \quad \mathbf{E}(\textit{dać}) = \{\text{np(nom)}, \text{np(dat)}\}$.

(20) $$\mathbf{F}(\textit{pożyczyć}) = \left\{ \begin{array}{ll} \{\text{np(nom)}, \text{np(acc)}\}, & \{\text{np(nom)}, \text{np(acc)}, \text{np(dat)}\}, \\ & \{\text{np(nom)}, \text{np(acc)}, \text{pp(od,gen)}\} \end{array} \right\} \text{(=to lend/borrow).}$$

(21) $$\mathbf{F}(\textit{powiedzieć}) = \left\{ \begin{array}{ll} \{\text{np(nom)}, \text{np(acc)}\}, & \{\text{np(nom)}, \text{np(acc)}, \text{np(dat)}\}, \\ \{\text{np(nom)}, \text{sentp(że)}\}, & \{\text{np(nom)}, \text{sentp(że)}, \text{np(dat)}\} \end{array} \right\} \text{(=to say).}$$

(22) $$\begin{aligned} \mathbf{L}^*(\textit{być}) = \{&\{\text{np(nom)}\}, \{\text{np(nom)}, \text{infp}\}, \{\text{np(nom)}, \text{advp}\}, \{\text{np(inst)}\}, \\ &\{\text{np(inst)}, \text{adjp(nom)}\}, \{\text{np(inst)}, \text{infp}\}, \{\text{np(inst)}, \text{advp}\}, \\ &\{\text{np(inst)}, \text{adjp(nom)}, \text{infp}\}, \{\text{np(inst)}, \text{adjp(nom)}, \text{advp}\}, \\ &\{\text{adjp(nom)}\}, \{\text{adjp(nom)}, \text{infp}\}, \{\text{adjp(nom)}, \text{advp}\}, \{\text{infp}\}, \{\text{advp}\}\}. \end{aligned}$$

(23) $\mathbf{L}^*(\textit{dać}) = \{\{\text{np(nom)}\}, \{\text{np(acc)}\}, \{\text{np(acc)}, \text{infp}\}, \{\text{np(dat)}\}, \{\text{infp}\}\}$.

(24) $\mathbf{E}^*(\textit{być}) = \{\{\text{np(nom)}, \text{advp}\}\}, \quad \mathbf{E}^*(\textit{dać}) = \{\{\text{np(nom)}\}, \{\text{np(acc)}, \text{infp}\}, \{\text{np(dat)}\})\}$.

(25) $\mathbf{L}^R(\textit{być}) = \{\{\text{np(nom)}\}, \{\text{np(nom)}, \text{advp}\}, \{\text{np(inst)}\}, \{\text{adjp(nom)}\}, \{\text{infp}\}, \{\text{advp}\}\}$.

Table 4: Toy examples of verb valences and various functions derived for them. We do not claim the completeness of valences given in the examples.

References

Bańko, Mirosław (ed.), 2000. *Inny słownik języka polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

Colmerauer, A., 1978. Metamorphosis grammar. In *Natural Language Communication with Computers*, Lecture Notes in Computer Science 63. New York: Springer, pages 133–189.

Durka, Piotr Jerzy, 2003. *Wstęp do współczesnej statystyki*. Warszawa: Adamantan.

Fast, Jakub and Adam Przepiórkowski, 2005. Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. In Zygmunt Vetulani (ed.), *Proceedings of the 2nd Language & Technology Conference, Poznań, Poland, April 21–23, 2005*. pages 191–195.

Halford, G. S., W. H. Wilson, and W. Phillips, 1998. Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral Brain Sciences*, 21(6):803–864.

Lepage, Yves, 2001. Analogy and formal langauges. In *Proceedings of Formal Grammar/Mathematics of Language Conference. August 10–12, 2001. Helsinki, Finland. Electronic Notes in Theoretical Computer Science, vol. 53*. Elsevier.

Mędak, Stanisław, 2005. *Praktyczny słownik łączliwości składniowej czasowników polskich*. Kraków: Universitas.

Mohri, Mehryar and Brian Roark, 2005. Structural zeros versus sampling zeros. Technical Report CSEE-05-003, OGI School of Science & Engineering, Oregon Health & Science University.

Polański, Kazimierz (ed.), 1980–1992. *Słownik syntaktyczno-generatywny czasowników polskich*. Wrocław / Kraków: Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN.

Przepiórkowski, Adam, 2006. What to acquire from corpora in automatic valence acquisition. In Violetta Koseska-Toszewa and Roman Roszko (eds.), *Semantyka a konfrontacja językowa (3)*. Warszawa: Slawistyczny Ośrodek Wydawniczy PAN.

Przepiórkowski, Adam and Jakub Fast, 2005. Baseline experiments in the extraction of Polish valence frames. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski (eds.), *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM'05 Conference held in Gdańsk, Poland, June 13–16, 2005*. New York: Springer, pages 511–520.

Świdziński, Marek, 1992. *Gramatyka formalna języka polskiego*. Wydawnictwa Uniwersytetu Warszawskiego.

Świdziński, Marek, 1994. Syntactic dictionary of Polish verbs. Uniwersytet Warszawski / Universiteit van Amsterdam.

Woliński, Marcin, 2004. *Komputerowa weryfikacja gramatyki Świdzińskiego*. Ph.D. thesis, Instytut Podstaw Informatyki PAN.