# Automatic acquisition of subcategorisation from large text corpora

Adam Przepiórkowski

INSTITUTE OF COMPUTER SCIENCE
POLISH ACADEMY OF SCIENCES
ul. J. K. Ordona 21, 01-237 Warszawa

Explorations in Syntactic Government and Subcategorisation
Cambridge, 2 September 2011





## Outline







## Valence

Valence (valency) of a predicate:

- combinatory potential of a predicate, i.e.,
- a list of its possible arguments and their features.

Understood:

- morphosyntactically — HELFEN: ⟨NP[NOM], NP[DAT]⟩
- semantically:
  - semantic roles — HELFEN: ⟨Agent, Patient⟩
  - semantic categories (selectional preferences). — HELFEN: ⟨+Anim, +Anim⟩

Near-synonyms, overlapping terms:

- valence frame,
- argument structure,
- subcategorisation.





## More terminology

Terminological confusion, especially regarding notions such as *argument* and *complement*.

Here:

- if it combines with a predicate, it's a *dependent*,
- ***dependents* = *arguments* + *adjuncts***,
- ***arguments* = *subject* + *complements***.

Other terms:

- *arguments*: *(inner/internal) participants*, confusingly: *complements*,
- *adjuncts*: *outer/external participants*, *(free) adverbials*, *circumstantials*, *modifiers*, incorrectly: *adverbs*.

Sometimes *complement* is used in the meaning of *dependent* (very confusing).

## Argument / adjunct dichotomy 1

Or rather: the great argument / adjunct hoax?

Various commonly assumed tests flawed **conceptually**, e.g.: *head-dependence* ("arguments occur with a narrower/fixed range of heads than adjuncts"):

- there is a class of accusative temporal "adjuncts" in various languages (incl. Polish),
- but they may also occur as objects, e.g.:
- Spędził godzinę w szpitalu.
  spend.PERF.PAST.3.SG.M hour.ACC in hospital
  'He spent an hour in the hospital.'
- Godzina spędzona w szpitalu...
  hour.NOM spent.PASS.NOM in hospital





## Argument / adjunct dichotomy 2

Others flawed **empirically**, e.g., case assignment (e.g., Joan Maling on Finnish and Korean, my own work on Polish), various substitution tests, incl. *do so*, etc., cf.:

- The formalisms are thus more aptly referred to as information- or constraint-based rather than unification-based, and we will *do so* here.
- Its cord was useless in effect, so I'd no trouble in its removal; on *doing so* I was dumbfounded by its unexpected contents. (Meijs, 1984)
- ...featuring people (like Woody Allen himself) who can't sing and can't dance, but *do so* anyway.
- John kicked Mary and Peter *did so* to Ann. (Miller, 1992)





## Argument / adjunct dichotomy 3

Importantly, it is not clear that any two tests for the argument / adjunct distinction fully correlate.

On the other hand, there are some *prototypical* properties:

| argument relations | adjunct relations |
|---|---|
| repeatable | non-repeatable |
| specific to some verbs | general |
| (sometimes) obligatory | (always?) optional |

**Zero hypothesis**: continuum from prototypical arguments to prototypical adjuncts (contrary to textbooks).





## Digression 1

*Obviously*, natural languages are not context-free (CF).

Very influential: Chomsky 1963, *Formal Properties of Grammars*.

1970s and early 1980s: textbooks state it as an *obvious* fact that natural languages are non-CF; some cite some complicated phenomenon to support this claim (but without any proof).

Pullum and Gazdar (1982) point out *obvious* empirical and formal flaws in Chomsky 1963, including:

- starring a construction unstarred in a later publication,
- the (implicit) false assumption that a language containing a non-CF must itself be non-CF (counterexample: $a^n b^n c^n \in a^* b^* c^*$),
- the false assumption that a certain artificial language is non-CF.

An advice to linguistic beginners: don't believe everything you read in textbooks!

## Valence acquisition in an ideal world

I went to Cambridge. The conference was great.

1. Partition the data into sentential clauses.
2. Identify verbs and argument phrases in each clause.
3. Record each verb-argument(s) combination as a valence frame for the verb.





## Valence acquisition in the real world

1. Partition the data into sentential clauses.
2. Identify verbs and argument phrases in each clause.
3. Filter attested verb-argument(s) combinations based on their statistical distribution.
4. Record each remaining verb-argument(s) combination as a valence frame for the verb.

**Problems**

- Morphosyntactic annotation.
- Automatic identification of arguments.
- Identification of clauses.

Practically always **two stages**:

- **linguistic**: collection of evidence,
- **statistical**: deciding which evidence is reliable.





## Valence acquisition: some history 1

Two seminal papers published in 1993:

- Brent 1993:
  - *linguistic stage*: collection of observations from unannotated texts (lexical clues: function words, pronouns),
  - *statistical stage*: binomial hypothesis test (BHT),
  - small set of valence frames, not easily generalisable;
- Manning 1993:
  - *linguistic stage*: tagging and shallow parsing (identification of NPs, PPs, etc.),
  - *statistical stage*: BHT (as in Brent 1993),
  - large (but still restricted) set of valence frames.

Automatic acquisition usually targets only *morphosyntactic* valence (but see below).





## Valence acquisition: some history 2

Also full parsing used at the linguistic stage, e.g.:

- Ersan and Charniak 1995, 1996: probabilistic context-free grammar (PCFG) parser,
- Briscoe and Carroll 1997: probabilistic unification-based parser; probably the largest valence extraction system (Carroll *et al.*, 1998; Korhonen, 2000, 2002; Korhonen and Preiss, 2003; Korhonen *et al.*, 2006; Preiss *et al.*, 2007).

Linguistic stage is important.

Statistical stage — apparently less so.

## Valence acquisition: some history 3

Very popular (since Brent 1993 and Manning 1993) statistical measure: Binomial Hypothesis Test.

Based on two assumptions:

- the probability $P_{-F}$ of the false observation of a verb with the frame $F$ depends on $F$, not on the verb (and may be estimated),
- if a verb actually has the frame $F$, than the observed co-occurrence probability will be higher than $P_{-F}$.

Suppose that a verb $v$, occurring in the corpus $n$ times, was *observed* $k$ times with $F$. **Is $F$ really one of its valence frames?**

Let's **assume** for a second that we observed $v$ with $F$ by **mistake**. If we have $P_{-F}$, we can **calculate the probability of observing** $v$ with $F$ that many times. **If** this probability is **very small** (the usual threshold is 0.05), **then** our assumption was wrong and we contend that $v$ **actually takes** $F$.





## Valence acquisition: some history 4

Binomial probability of observing $F$ *exactly* $k$ times:

$$b(k; n, P_{-F}) = \binom{n}{k} P_{-F}^{k}(1 - P_{-F})^{(n-k)}$$

Binomial probability of observing $F$ *at least* $k$ times:

$$\sum_{i=k}^{n} b(i; n, P_{-F})$$

Method used also, e.g., in Ersan and Charniak 1995, 1996, Briscoe and Carroll 1997, Carroll and Rooth 1998, and Chesley and Salmon-Alt 2006.





## Valence acquisition: some history 5

Experiments in Fast and Przepiórkowski 2005:

- BHT overall much better than other statistical tests ($t$ test, log likelihood ratio, etc.).

But, in fact, various authors show that much simpler cut-off heuristics give similar results:

- e.g., Briscoe and Carroll 1997: add frame to the dictionary if it is supported by at least 10 observations,
- some other works using such heuristics: Kawahara *et al.* 2000, Korhonen 2002, Schulte im Walde 2002 and Dębowski 2009.





## Valence acquisition: some history 6

Automatic acquisition usually targets only *morphosyntactic* valence, but:

- morphosyntactic valence is evidence for alternations,
- alternations correlate highly with the semantic class of verbs (Levin, 1993),
- semantic classes of verbs determine their *semantic valence*.

Work in that direction: various papers and Ph.D. dissertation of Sabine Schulte im Walde, but also McCarthy 2001, Lapata and Brew 2004, etc.

Also, semantic valence may be read off semantic treebanks, e.g.:

- Gildea and Jurafsky 2002: using FrameNet/BNC,
- Gildea and Palmer 2002 and Surdeanu *et al.* 2003: using PropBank.

## Experiments for Polish: Aims

Aim:

- extraction of full valence frames (no distinction between arguments and adjuncts),
- not predefined,
- with frequencies,
- only morphosyntactic information.

Examples of extracted frames:

- `np(nom), np(acc)`
- `np(nom), nad+np(inst)`
- `np(nom), sie, inf`
- `np(nom), np(dat), ZE`
- `np(nom), sie, np(dat), z+np(inst)`

Empirical material: the IPI PAN Corpus of Polish (`http://korpus.pl`; Przepiórkowski 2004).





## Digression 2

A shameless plug...

The National Corpus of Polish:

- 1.5 billion ($1.5 \times 10^9$) words,
- annotated at various levels,
- with two different search engines,
- with a 1-million-word manually annotated corpus available on GNU GPL.

`http://nkjp.pl/`





## Experiments for Polish: Linguistic stage (and acronyms)

Linguistic stage — two methods in parallel:

- deep parsing with Świgra (Woliński, 2005):
  - implementation of Marek Świdziński's methamorphosis grammar of Polish,
  - `http://nlp.ipipan.waw.pl/~wolinski/swigra/`,
- shallow parsing with Spejd (Przepiórkowski and Buczyński, 2007):
  - Shallow Parser and Disambiguation Engine (SPADE; reserved: Soricut and Marcu 2003),
  - Shallow Parsing and Eminently Judicious Disambiguation,
  - Syntaktisches Parsing Entwicklungsystem Jedoch mit Disambiguierung,
  - Składniowy Parser (Ewidentnie Jednocześnie Dezambiguator),
  - `http://nlp.ipipan.waw.pl/Spejd/`,
  - small grammar written specifically for this task.





## Experiments for Polish: Świgra

Świgra:

- bottom-up chart-like constituency parser,
- implementing the largest grammar for Polish,
- almost 500 DCG-like rules,
- *no* valence dictionary,
- so each verb assumed to have all possible valence frames,
- *not* probabilistic,
- so parses not ranked.

An instance of the Expectation Maximisation (EM) algorithm used to pick one parse for each sentence (72.6% correctness; Dębowski 2009).

## Experiments for Polish: Spejd

Spejd:

- a cascade of rules,
- each rule specifies a regular grammar,
- grouping and disambiguations on strings found,
- no explicit representation of ambiguities.

A Spejd grammar:

- written within a few weeks,
- around 350 different rules (some repeated), incl.:
  - around 130 morphological rules,
  - around 100 rules identifying simple syntactic words,
  - around 120 rules identifying NPs, PPs, etc.,
- available at `http://nlp.ipipan.waw.pl/PPJP/`.





## Experiments for Polish: Statistical stage 1

Interesting idea (Dębowski and Woliński, 2007): split valence frame observations into argument observations:

- if an argument co-occurs with a verb *very* often, it must occur in each valence frame of the verb,
- if it co-occurs *very* rarely, it cannot occur within any valence frame of the verb,
- parameters "very often" and "very early" learned for each argument type on the basis of an existing small valence dictionary (parameters independent of verbs),
- also, a table learned, specifying whether:
  - two argument types always co-occur,
  - one implies the other,
  - they never co-occur,
  - they are independent.





## Experiments for Polish: Statistical stage 2

In effect:

- frame observations decomposed into argument observations,
- arguments learned,
- frames recomposed from arguments (and their co-occurrence table).

See Dębowski 2009 (in *Language Resources and Evaluation*) for details.

OK, so we built two valence dictionaries (one using Świgra parsing, the other using Spejd). How good and useful are they?





## Evaluation: introduction

As usual, evaluation may be:

- extrinsic (*in vivo*): evaluate an application (parser) making use of the dictionary,
- intrinsic (*in vitro*): evaluate the dictionary alone.

Very few attempts at the extrinsic evaluation:

- Briscoe and Carroll 1997 show that automatically acquired valence dictionary improves parsing results,
- Carroll *et al.* 1998 show that the improvement is statistically significant,
- also Kawahara *et al.* 2000, Carroll and Fang 2004 and Arun and Keller 2005.

## Dictionary-based evaluation of valence dictionaries

Two kinds of intrinsic evaluation:

- dictionary-based: compare results to a manually created valence dictionary,
- corpus-based: compare results to a corpus annotated with valence frames.

Dictionary-based evaluation:

- cheap (if a gold-standard valence dictionary exists),
- measures *type* recall and *type* precision,
- i.e., frequent and rare frames carry same weight;
- both measures lower than "in reality" (Manning, 1993; Briscoe and Carroll, 1997; Preiss *et al.*, 2007):
  - some theoretically possible frames don't occur in corpora,
  - not all frames occurring in corpora listed in dictionaries;
- valence dictionaries may differ substantially (Przepiórkowski and Fast, 2005).





## Corpus-based evaluation of valence dictionaries

Corpus-based evaluation:

- annotate each verb in a corpus with the valence frame of that occurrence of the verb,
- measure *token* recall,
- if all occurrences of a number of verbs in a corpus so annotated, also some estimation of type precision and type recall (Briscoe and Carroll, 1997).

Much better approximation of the real usefulness of the dictionary.





## Evaluation for Polish

There are three manually constructed valence dictionaries:

- Polański 1992,
- Świdziński 1998,
- Bańko 2000.

For **dictionary-based** evaluation, a sample of 200 verbs selected, a majority voting (MV) dictionary constructed. (An interpretive process.)

For **corpus-based** evaluation, 12 verbs of different frequencies selected (4 very frequent, 4 of medium frequency, 4 rare). About 100 occurrences of each, randomly selected from a balanced subcorpus, were annotated.





## Evaluation results

Summary of the results for the initial valence dictionary (cf. Przepiórkowski 2009, 2008 for details):

| | | baseline | Świgra | Spejd |
|---|---|---|---|---|
| dictionary based evaluation | P | 47.41 | **54.18** | 53.01 |
| | R | 15.15 | **35.55** | 31.45 |
| | F | 22.96 | **42.93** | 39.48 |
| corpus-based evaluation | | 10.25 | 26.80 | **32.28** |

The best baseline: assign each verb two frames:

- transitive: `np(nom), np(acc)`,
- intransitive: `np(nom)`.

## Results: examples

```
'gadać' => {
 'np(nom)' => 58,
 'np(acc),np(nom)' => 8,
 'np(nom),PZ' => 5
}
'gasić' => {
 'np(nom)' => 12,
 'np(acc),np(nom)' => 5
}
'gasnąć' => {
 'np(nom)' => 10,
 'nad+np(inst),np(nom)' => 3
}
'generować' => {
 'np(nom)' => 12,
 'np(acc),np(nom)' => 6
}
```

```
'ginąć' => {
 'np(nom)' => 49
}
'gniewać' => {
 'np(nom),sie' => 22,
 'np(nom),sie,ZE' => 3
}
'godzić' => {
 'na+np(acc),np(nom),sie' => 42,
 'inf,np(nom),sie' => 28,
 'np(nom),w+np(acc)' => 26,
 'np(nom),sie' => 17,
 'np(acc),np(nom),w+np(acc)' => 4,
 'np(acc),np(nom)' => 4,
 'np(acc),np(nom),sie' => 4,
 'np(nom)' => 3
}
```





## Conclusions

Reasonable evaluation of automatic valence extraction — still an open issue.

We need more and better resources (coming into being for Polish now).

But, in general:

- valence may be acquired automatically,
- quality still not too good,
- but already helpful for automatic parsing,
- and, hopefully, for linguistic **explorations in syntactic government and subcategorisation**.

**Thank you for your attention!**






ACL (2003). *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo.

Arun, A. and Keller, F. (2005). Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 306–313, Ann Arbor, MI.

Bańko, M., editor (2000). *Inny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, **19**(2), 243–262.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Applied Natural Language Processing Conference*, pages 356–363, Washington, D.C. ACL.

Carroll, G. and Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd conference on empirical methods in natural language processing (EMNLP 3)*.

Carroll, J. and Fang, A. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 107–114, Sanya City, China.

Carroll, J., Minnen, G., and Briscoe, T. (1998). Can subcategorisation probabilities help a statistical parser? In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 118–126, Montreal.

Chesley, P. and Salmon-Alt, S. (2006). Automatic extraction of subcategorization frames for French. In LREC (2006), pages 253–258.

Chomsky, N. (1963). Formal properties of grammars. In R. D. Luce, R. R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume II. John Wiley, New York.

Dębowski, Ł. (2009). Valence extraction using the EM selection and co-occurrence matrices. arXiv:0711.4475v4 [cs.CL] 10 Dec 2008. To appear in *Language Resources and Evaluation*.

Dębowski, Ł. and Woliński, M. (2007). Argument co-occurence matrix as a description of verb valence. In Vetulani (2007), pages 260–264.

Ersan, M. and Charniak, E. (1995). A statistical syntactic disambiguation program and what it learns. Technical Report CS-95-29, Brown University.

Ersan, M. and Charniak, E. (1996). A statistical syntactic disambiguation program and what it learns. In S. Wermter, E. Riloff,

and G. Scheler, editors, *Connectionist, Statistical and Symbolic Approaches in Learning for Natural Language Processing*, pages 146–157. Springer-Verlag, Berlin.

Fast, J. and Przepiórkowski, A. (2005). Automatic extraction of Polish verb subcategorization: An evaluation of common statistics. In Z. Vetulani, editor, *Proceedings of the 2nd Language & Technology Conference*, pages 191–195, Poznań, Poland.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, **28**(3), 245–288.

Gildea, D. and Palmer, M. (2002). The necessity of parsing for predicate argument recognition. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, PA.

Kawahara, D., Kaji, N., and Kurohashi, S. (2000). Japanese case structure analysis by unsupervised construction of a case frame dictionary. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 432–438, Saarbrücken.

Kim, S. and Maling, J. (1993). Syntactic case and frequency adverbials in Korean. In *Harvard Studies in Korean Linguistics V*, pages 368–378.





Kim, S. and Maling, J. (1996). Case assignment in the *siphta* construction and its implications for case on adverbials. In R. King, editor, *Description and Explanation in Korean Linguistics*, pages 141–179. Cornell University Press.

Korhonen, A. (2002). *Subcategorization Acquisition*. Ph. D. dissertation, University of Cambridge.

Korhonen, A. and Preiss, J. (2003). Improving subcategorization acquisition using word sense disambiguation. In ACL (2003).

Korhonen, A., Krymolowski, Y., and Briscoe, T. (2006). A large subcategorization lexicon for natural language processing applications. In LREC (2006), pages 1015–1020.

Korhonen, A.-L. (2000). Using semantically motivated estimates to help subcategorization acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. ACL.

Lapata, M. and Brew, C. (2004). Verb class disambiguation using informative priors. *Computational Linguistics*, **30**(2), 45–73.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.





LREC (2006). *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, Genoa. ELRA.

Maling, J. (1989). Adverbials and structural case in Korean. In S. Kuno, I.-H. Lee, J. Whitman, S.-Y. Bak, Y.-S. Kang, and Y.-j. Kim, editors, *Harvard Studies in Korean Linguistics III*, pages 297–308, Cambridge, MA.

Maling, J. (1993). Of nominative and accusative: The hierarchical assignment of grammatical case in Finnish. In A. Holmberg and U. Nikanne, editors, *Case and Other Functional Categories in Finnish Syntax*, pages 51–76. Mouton, Dordrecht.

Manning, C. D. (1993). Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, OH.

McCarthy, D. (2001). *Lexical Acqusition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph. D. dissertation, University of Sussex.

Meijs, W. (1984). 'You can do so if you want to': Some elliptic structures in Brown and LOB and their syntactic description. In J. Aarts and W. Meijs, editors, *Corpus Linguistics: Recent*





*Developments in the Use of Computer Corpora in English Language Research*, pages 141–162. Rodopi, Amsterdam.

Miller, P. H. (1992). *Clitics and Constituents in Phrase Structure Grammar*. Garland, New York.

Polański, K., editor (1980–1992). *Słownik syntaktyczno-generatywny czasowników polskich*. Zakład Narodowy im. Ossolińskich / Instytut Języka Polskiego PAN, Wrocław / Cracow.

Preiss, J., Briscoe, T., and Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 912–919, Prague.

Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Przepiórkowski, A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Przepiórkowski, A. (2009). Towards the automatic acquisition of a valence dictionary for Polish. In M. Marciniak and A. Mykowiecka, editors, *Aspects of Natural Language Processing*, volume 5070 of *Lecture Notes in Computer Science*, pages 191–210. Springer-Verlag, Berlin.

Przepiórkowski, A. and Buczyński, A. (2007). ♠: Shallow Parsing and Disambiguation Engine. In Vetulani (2007), pages 340–344.

Przepiórkowski, A. and Fast, J. (2005). Baseline experiments in the extraction of Polish valence frames. In M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, editors, *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, pages 511–520. Springer-Verlag, Berlin.

Pullum, G. K. and Gazdar, G. (1982). Natural languages and context-free languages. *Linguistics and Philosophy*, **4**, 471–504.

Schulte im Walde, S. (2002). A subcategorisation lexicon for German verbs induced from a lexicalised PCFG. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, pages 1351–1357, Las Palmas. ELRA.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT-NAACL 2003*, pages 149–156, Edmonton. ACL.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In ACL (2003).





Vetulani, Z., editor (2007). *Proceedings of the 3rd Language & Technology Conference*, Poznań, Poland.

Woliński, M. (2005). An efficient implementation of a large grammar of Polish. *Archives of Control Sciences*, **15**(3), 251–258.

Świdziński, M. (1998). Negacja w polszczyźnie: uwikłania składniowe imiesłowów, gerundiów i quasi-gerundiów. *Prace Filologiczne*, **XLIII**, 411–420.