PolVerbNet: an experimental database for Polish verbs

Barbara Gawronska

University of Skövde
Dept. of Languages
Box 408, S-541 28 Skövde, Sweden
barbara.gawronska@isp.his.se

Abstract

The semantics of verbs implies, as is known, a great number of difficulties, when it is to be represented in a computational lexicon. The Slavic languages are especially challenging in respect of this task because of the huge complexity of verbs, where the stems are combined with prefixes indicating aspect and Aktionsart. The current paper describes an approach to build PolVerbNet, a database for Polish verbs, considering the internal structure of the aspect-Aktionsart system. PolVerbNet is thus implemented in a larger English-Polish MT-system, which incorporates WordNet. We report our translation procedure and the system's performance is evaluated and discussed.

Keywords

verbs, WordNet, aspect, Aktionsart, Polish

Introduction: Aspect and Aktionsart in Machine Translation

Commercial MT-systems of today are mostly quite good at handling word order, morphological agreement within noun phrases and the choice of grammatical case forms. Aspect and Aktionsart remain a problem. When testing an on-line English-Polish translation system (www.poltran.com) we achieved very poor results with regard to the choice of Aktionsart (only 21% recall and ca 28% precision for an input consisting of 80 sentences). Most of the output was incomprehensible, although other features of POLTRAN indicate that the system includes a great deal of linguistic knowledge.

There is a gap between practical MT-applications for Slavic languages and the theoretical work on aspect and Aktionsart (enumerating all relevant references would fill several pages; thus, we only mention some work related to computational linguistics, like Reichenbach, 1947; Johnson, 1981; Maegaard, 1982; van Eynde et al, 1985; Dowty, 1986; Santos, 1992). In implementation-oriented computational research, most effort has been made with tense and aspect systems in Romance and Asian languages (Han et al., 2000; Murata et al., 1999; Palmer & Wu, 1995; Dorr et al., 2000; Santos,1992).

The notions "aspect" and "Aktionsart" have been defined in many different ways. Although there is no general agreement as to the details in the proposed definitions, most authors characterize the distinction between the perfective and the imperfective aspect as the difference between events that are "countable", "bounded", "unique", "constant" (perfective), and those that are "unbounded", with no clear temporal limits (imperfective). The term Aktionsart is most frequently used when referring to the internal spatiotemporal structure of the event. Certain authors (e.g. Comrie 1976, Langacker 1982) include the notion of Aktionsart into the one of aspect. Furthermore, there is no agreement as to the question of universality of the two categories. The most plausible solution to the universality problem seems to be that all languages are able to express "an aspectual perspective" (Nakhimovsky 1987), and that all languages make use of different Aktionsarts, but the overt markers of these categories are different for different language groups. Even languages belonging to the same family display certain differences as to the internal structure of the aspect/Aktionsart systems (see figures 1a-1c in the next section). Thus, an analysis of an aspectual system in Korean or Spanish is not directly applicable in a translation system for Slavic languages (although there can be certain similarities between the different aspectual systems). Establishing language-specific correspondence links between the inherent semantics of verbal stems, the possible syntactic and semantic alternations of the verb, and the contextual markers that indicate aspectual perspective and Aktionsart is a necessary prerequisite for correct choice of the translation equivalent (van Eynde et al, 1985; Levin, 1993; Sanford Pedersen, 1999; Pustejovsky, 1991). This fact is not taken into account in most commercial MT-systems for Slavic. The lexical information is mostly copied from traditional bilingual dictionaries, and the worst errors are handled by ad-hoc rules, or, at best, by statistical methods.

In the following, we try to extract common denominators from previous work on aspect/Aktionsart and try to utilise them in connection to a well-known lexical database, WordNet (Miller, 1990,1995; Vossen, 1998), in order to show that machine translation between Polish and English can be improved.

The Slavic verbs

The semantics of verbs, including the semantics of morphological categories marked on verbs, is, in general, difficult to represent in a consistent way in a computational lexicon. In the Slavic languages, this task becomes extremely complicated because of the enormous number of complex verbs, where stems are combined with prefixes expressing either aspect, or Aktionsart, or both. It is commonly known that the Slavic verbs occur either in the perfective or the imperfective form; however, the frequently used term aspectual pairs "should in many cases be replaced with the term aspectual clusters" (Gawronska 1993:77). It is possible to derive several perfective forms from one Slavic imperfective verb. For example, the Polish verb *kończyć* (*finish*, imperfective) is related to *skończyć, zakończyć, ukończyć, dokończyć* – perfective forms that differ as to their Aktionsart. Furthermore, the semantics of the verbal stems in Slavic has in some cases undergone historical changes that affected the aspect clusters. A good example is provided by the Old Slavic stems **legeti* (lay) and **kladt* (put) (Cyganenko, 1970): **legeti* (*lay*; from Indoeuropean *legti*, cf. German legen, lagern, Latin lectus) is the common ancestor of Russian *ložit*, Polish *łożyć*, Bulgarian *legna*, and Czech *ložiti*. However, only Bulgarian has retained original aspectual pairs with the stem *leg-* (e.g. *slagam/sloza* – 'put'). In contemporary Czech, Polish and Russian, the verbs originating from the stem *kladti* (put; Ru. *klast'*, Pol. *kłaść*, Cz. *klasti*) function as imperfective counterparts of the *leg-* formations with the meaning 'put'.

Figures 1a-1c show the central parts of these aspectual clusters; the verbs are provided with their most frequent English translation equivalents. The kernel of each cluster is a morphologically underived imperfective verb. The differences between the kernel verb and the derived verbs can be classified as follows:

- Aspect difference only: for example, *kłaść/położyć* can appear in the same context and refer to the same type of action, the distinction being very similar to the one between simple tenses and continuous tenses in English (e.g. *Kładł książki na stole – He was putting the books on the table* vs. *Położył książki na stole – He put the books on the table*).
- Aktionsart difference, i.e. a difference concerning the spatiotemperal structure of the event, especially the starting point and the end point of an action. E.g. *kłaść* and *dokładać* are both imperfective verbs, but *kłaść* refers to the action of putting in general; the *dokładać* is more specific: it expresses the action of adding, i.e. putting an object into a place where similar objects are already present.
- Aspect and Aktionsart difference: e.g. *kłaść* ('put', imperfective)/ *złożyć* ('put together',perfective); here, both the value of aspect and the internal structure of the event are affected.
- 'Metaphorical shift': some of the verbs in the 'put' cluster have undergone a semantic shift from verbs of physical manipulation to communication verbs (similar semantic shifts in English verbs are described in Sweetser, 1987). E.g. *wykładać* retains its original meaning ('put out/take out'), but it also *means* 'lecture'. Metaphorical shifts may be combined with aspect and/or Aktionsart difference.

Figure 1 does not contain reflexive verb forms; we also excluded forms with double prefixes (*ponakładać, pozakładać* etc). Reality is rather more complex than the picture below.

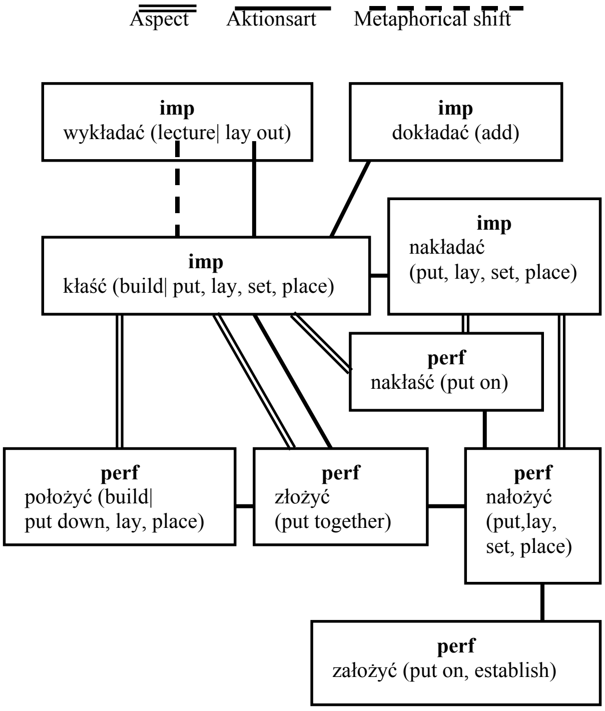


Figure 1a: A fragment of a Polish aspect-Aktionsart network

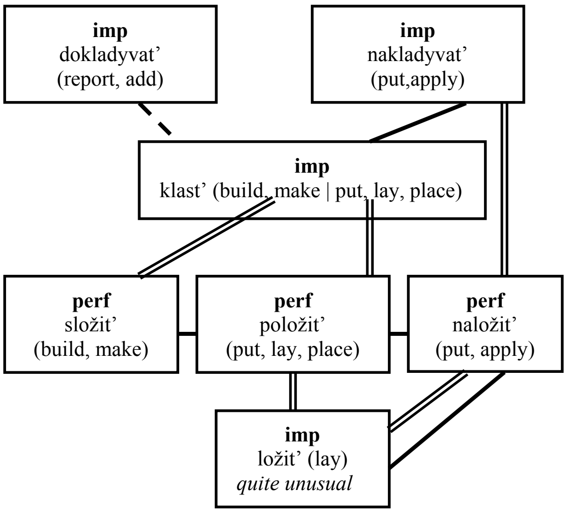


Figure 1b: A fragment of a Russian aspect-Aktionsart network

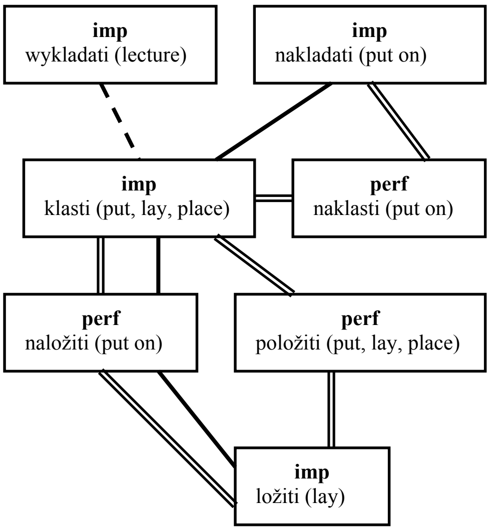


Figure 1b: A fragment of a Czech aspect-Aktionsart network

The goal of the current project: a Polish VerbNet

The project referred to here has emerged from a domain-restricted English-Swedish-Polish translation system dealing with weather forecasts (Gawronska & Duczak, 2000). Our attempts to choose the correct Aktionsart of the few Polish verbs occurring within this particular domain were quite encouraging, so we decided to develop the model for a larger domain. The early system for translation of weather forecasts utilised a very restricted lexical database, designed and implemented by our research group. The next step towards a system with a larger coverage was to incorporate WordNet (Miller, 1995; Vossen, 1998) into the system.

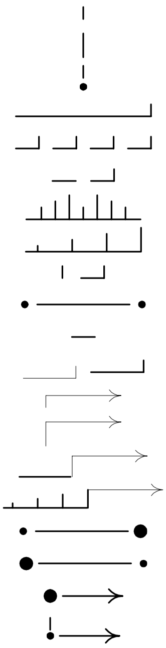
We considered it important not to force the Polish verbs directly into the semantic network designed for the English verbs. We organised the Polish verbs into a separate network, where relations between imperfective underived verbs and imperfective/perfective complex verbs are treated as a special kind of troponymy (a transitive relation similar to hyponymy: X is a troponym of Y if X means ‘doing Y in a specific way’ – Miller 1990, 1995). This approach seems to be linguistically more relevant than attaching all possible Polish translation equivalents of an English verb to one “synset” node in the English lexical network. Furthermore, the morphological markers of aspect and Aktionsart are productive. Any foreign verb that enters the Polish lexical system develops into an aspectual cluster (e.g. a newcomer like the English verb *click* may be realised as *kliknąć* (perf), *klikać* (imperf), *poklikać* (imp+distributive) etc). Thus, the lexical model should be prepared for the introduction of new verb groups, not of single loan verbs only. Finally, if the Polish „VerbNet“ functioned satisfactorily in translation, the model could easily be adopted for organising lexicons for other Slavic languages.

The lexical organisation of Polish aspect and Aktionsart clusters

The model of lexical organisation of the Polish verbs we implemented is based on the following theoretical claims:

- The imperfective aspect expresses the fact that the reference time is properly included in the event time; this means that an imperfective event is normally seen “from inside” (Johnson, 1981; Maegaard, 1982 and van Eynde et al., 1985)
- There is a parallel between the distinction bounded/unbounded, or count/mass, and the perfective/imperfective distinction (Papprotté, 1988; Talmy, 1988); a perfective event in Slavic is seen as a constant, bounded whole
- Aktionsart in Polish is a semantic category which characterizes the internal spatiotemporal structure of an event and/or the relations between the event itself, its cause and/or its result
- A Polish verb prefix can in combination with different verb stems express either a change of Aktionsart (i.e., in most cases, a change of the internal spatiotemporal structure of the event), or only a change of aspect, or it can express an aspect shift and an Aktionsart shift at the same time. The lexical model must take these three possible functions of the prefixes into account (Agrell, 1908).

The last mentioned work from the beginning of the 20th century provides an outstanding analysis of the Polish verb prefixes. An interesting fact is that Agrell’s way of representing the semantics of Aktionsart displays a considerable similarity to the representations used by cognitive linguists of today (Lakoff, 1993; Jackendoff, 1983; Langacker, 1991). Agrell’s Aktionsart-schemata are shown in Figure 2 (the picture copied from the original paper). The horizontal lines symbolize the temporal dimension of the event; the vertical lines represent result achievement. Points stand for the beginning and the end of a spatiotemporal trajectory; when in focus, these ends are marked by larger points. Arrows indicate that the time before or after the event is of importance for the particular Aktionsart.



Aktionsart.	Schema	Typverbum.
1) resultativ		*zemrzeć* (ev. *skończyć*)
2) effektiv		*wykończyć*
3) momentan		*ukluć* (ev. *skończyć*)
4) durativ		*ukończyć*
5) distributiv		*pokończyć*
6) final		*dokończyć*
7) akkurativ		*odróżnić* (*odegrać*)
8) augmentativ		*rozróznić* (*rozdrażnić*)
9) majorativ		*podrożeć*
10) perdurativ		*przenocować*
11) präteritiv		*poszukać*
12) konsekutiv		*pochwalić*
13) definitiv		*zakonczyć*
14) eff.-definitiv		*przebudzić*
15) dur.-definitiv		*nakierować*
16) aug.definitiv		*wzbudzić*
17) terminativ		*przyjść*
18) perkursiv		*zajść*
19) kursiv		*pójść*
20) inchoativ		*zagrać*

Figure 2: Agrell's schemata for Polish verbs

The structure of the experimental Polish VerbNet is based on Agrell's schemata combined with the above-enumerated theoretical assumptions about the perfective/imperfective distinction. The schemata shown in figure 2 were enriched by a more detailed specification of the semantics of trajectors (TR; objects that move or are in focus) and landmarks (LM; 'stable' objects, in relation to which the trajectors and their movements are regarded). In Figure 3, we show the results of our analysis

of the "put"-cluster in Polish.

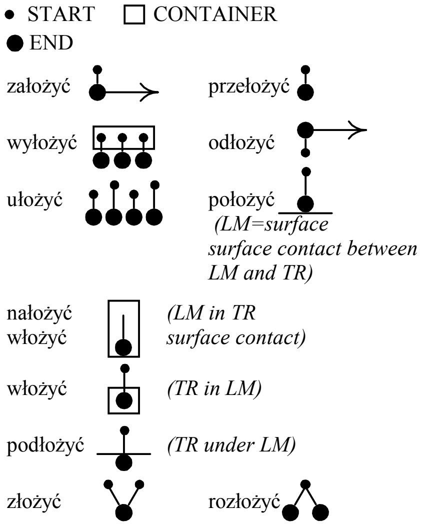


Figure 3: The analysis of the "put"-cluster in Polish

The database has been structured in accordance with the Agrell-inspired analysis. In Table 1, we show a fragment of the 'put'-cluster, as implemented in the data base.

Verbs belonging to a lexical cluster are organized under a top node – normally occupied by an underived imperfective verb. The top node is connected to the identification number of its English equivalent in WordNet (for the reason of space, we do not show these connections in Table 1). Each verb can be specified with regard to the features of its trajector, the features of the trajectory (e.g. in the case of *włożyć*, the trajector is outside the landmark at the beginning of the action, an inside the landmark when the action ends) and a feature kalled "plexity" (iterative and distributive events are treated as multiplex, punctual – as uniplex). Verbs on lover levels in the hierarchy inherit the semantic features of the top-node verb, aspect excluded. The inheritance process goes on along links called 'pointers'. Pure aspect change is treated as a change of the feature "boundedness" (presence or absence of temporal limits) and, when both involved verbs are morphologically derived, even as a change of "plexity" All imperfective verbs are considered as unbounded, and all perfective as bounded. When an imperfective form is created from a derived perfective, uniplex events may become multiplex, e.g. *odłożyć* ('put aside', perfective – *odkładać* ('put aside', iterative, imperfective). As mentioned in section 2, some of the verbs in the "put"(*kłaść*)-cluster have undergone a metaphorical shift from verbs of physical manipulation to communication verbs. The verbs *przekładać/przełożyć* and *wykładać* can still refer to physical manipulations ('put from one place to another; shift' respectively 'put out/take out'), but are also used as equivalents of *translate* and *lecture*, respectively. The latter senses are linked to the 'put'-cluster by connections ("pointers") labelled 'metaphorical shift' (in Table 1, pointer 1050 belongs to the 'metaphorical shift' markers). The 'communication' senses inherit the general characteristics of the image-schema of the verb of physical object manipulation, but they do not inherit the exact characteristics of the trajector. In Table 1, the 'lecture' sense of *wykładać*, labelled 1050, does not inherit the trajector features 'solid' and 'multiplex' from the non-metaphorical verbs. It is instead provided provided with a separate trajector specification (trajector = scientific area). In this way, metaphors like "minds are containers" and "messages are objects" can be inferred from the database.

Because of the inheritance mechanism, several members of an aspect and Aktionsart cluster can be connected to the same translation equivalent in the English WordNet. A single English sentence can contain the following cues as to the choice of the right Polish verb form: preposition or particles, tense, temporal and spatial adverb and semantic features of the complements of the verb.

Figure 4 shows how these cues are utilized when choosing translation equivalents.

Polish ID	Verb	Aspect	Trajector	Start	End	Plexity	Pointers
1000	kłaść	imp	solid	not(contact(tr,lm))	contact(tr,lm)	uni	
1032	włożyć	perf		outside(tr,lm)	in(tr,lm)	uni	1033
1033	wkładać	imp				multi	
1045	wyłożyć	perf	multi	in(tr,lm)	out_off(tr,lm)	uni	1046
1046	wykładać	imp				multi	1050met
1050met	wykładać	imp	scientific domain				

Table 1: A fragment of the implemented PolVerbNet

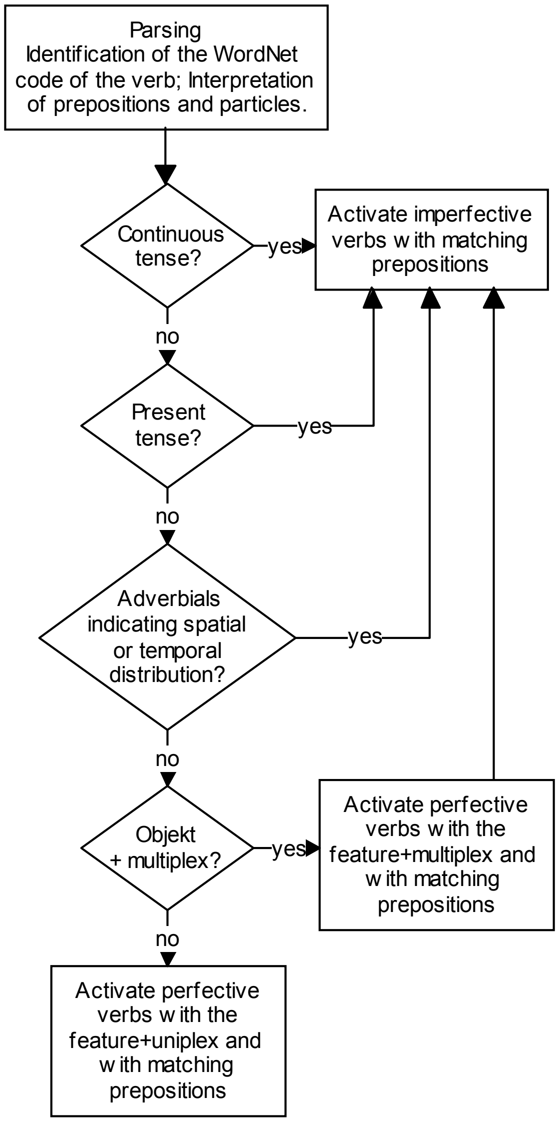

Parsing
Identification of the WordNet code of the verb; Interpretation of prepositions and particles.
Continuous tense?
yes
Activate imperfective verbs w ith matching prepositions
no
Present tense?
yes
no
Adverbials indicating spatial or temporal distribution?
yes
no
Objekt + multiplex?
yes
Activate perfective verbs w ith the feature+multiplex and w ith matching prepositions
no
Activate perfective verbs w ith the feature+uniplex and w ith matching prepositions

Figure 4: Translation procedure

Evaluation and implications for further research

The currently implemented database contains Polish verbs of motion and appearance, and Polish equivalents of English 'verbs of putting' and 'verbs of pouring' (according to Levin's (1993) classification). Many members of those categories had evolved into verbs of mental activity and/or verbal communication (Sweetser 1987). Thus, many verbs in the database belong to multiple troponymy hierarchies: some verbs are – pragmatically – troponyms of e.g. 'communicate' and, at the same time, they are – on a deep semantic level – troponyms of the top-node verb in a cluster of morphologicaly related verbs with different aspect and Aktionsart values.

We tested the performance of the algorithm shown in Figure 4 with respect to the experimental database by translating 200 English sentences into Polish. The sentences were taken from news reports and from dictionary examples. The test material contained eight English verbs (*put, pour, stream, flow, move, come, appear and follow*), modified by different particles or prepositions. We achieved ca 83% recall at 85% precision, which is a clear improvement compared with the commercial system mentioned in the introduction. However, the results would not achieve this degree of over 80% precision and recall if we tried to translate coherent texts, since aspect also has a foreground/background marking function in discourse (Dowty, 1986; Paprotté, 1988; Hopper & Thompson 1980). Nevertheless we can conclude that preserving the language-specific semantic and morphological relations in the database leads to better results than attaching all possible translation equivalents directly to nodes in a lexical network created for a structurally different language.

Most of the wrong translations depended actually on difficulties in choice of the right WordNet 'synset' when parsing the English input. One of our future goals is thus further work on modifying the current version of WordNet to achieve a higher disambiguation rate in parsing (some of the work, inspired by, among others, Resnik, 1992 and Leacock et al., 1998, is already implemented). Furthermore, our plans include theoretical research on discourse-related cues for Aspect choice, implementing more verb clusters and testing the system for translation of coherent texts.

References

Agrell, S. (1908). Aspektänderung und Aktionsartbildung beim polnischen Zeitworte. Lund (diss).

Cyganenko, G. P. (1970). Etimologičeskij slovar' russkogo jazyka. Kiev, Izdatel'stvo "Radjans'ka škola".

Comrie, B. 1976. Aspect. Cambridge: Cambridge University Press.

Dang, H. T., Kipper, K., Palmer, M., Rosenzweig, J. (1998). Investigation Regular Sense Extensions Based on Intersective Levin. In ACL/COLING 98, Proceedings of the 36th annual meeting of the Association for Computational Linguistics (pp. 293-299).

Dorr, B. J., Garman, J., Weinberg, A. (1995). From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT. Machine Translation 9 (pp. 221--250).

Dorr, B. J., Levov, G-A., Lin, D. (2000). Building a Chinese-English Mapping between Verb Concepts for Multilingual Applications. In Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000 (pp. 1--12). Springer-Verlag.

Dowty, D.R. (1986). The effects of aspectual class on the temporal structure of discourse: Semantics or Pragmatics? Linguistics and Philosohy 9, 37--61.

Eynde, F. van (1988). The analysis of tense and aspect in EUROTRA. In Proceedings of Coling 88 (pp. 699--704). Budapest.

Fillmore, C. & Atkins, B. (1992). Toward a Frame-based Lexicon: The Semantics of RISK and its Neighbours. In A. Lehrer & E. Kittay (Eds.), Frames, Fields, and Contrasts (pp. 75--102). Hillsdale, NJ:Lawrence Erlbaum.

Gawronska, B. (1993). An MT oriented Model of Aspect and Article Semantics. Lund: Lund University Press.

Gawronska, B. & Duczak, H. (2000). Understanding Politics by Studying Weather. In: White, J.S. (ed.): Envisioning Machine Translation in the Information Future 147-157. Berlin/ New York: Springer.

Han, C., Lavoie, B., Palmer, M., Rambow, O., Kittredge, R., Korelsky, T., Kim, N., Kim, M. (2000). Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System. In Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000 (pp. 40--53). Springer-Verlag.

Hopper, J.P. & Thompson, S. (1980). Transitivity in grammar and discourse. Language 56:2, 251--299.

Jackendoff, R.S. (1983). Semantics and Cognition. Cambridge, MA: MIT Press.

Johnson, M. (1981). A unified temporal theory of tense and aspect. In P. Tedeschi & A. Zaenen (Eds.), Syntax and Semantics, vol. 14: Tense and Aspect. New York: Academic Press.

Lakoff, G. (1993). The Contemporary Theory of Metaphor. In Ortony, A. (Ed.), Metaphor and Thought, 2d ed. Cambridge: Cambridge University Press.

Langacker, R. (1991). Concept, Image, and Symbol. The Cognitive Basis of Grammar. Berlin/New York: Mouton de Gruyter.

Leacock, C., Miller, G.A., Chodorow, M. (1998). Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, 24/1, 147--165.

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

Maegaard, B. (1982). The transfer of finite forms in a machine translation system. In Proceedings of Coling 82 (pp. 190--194), Prague.

Miller, G.A. (Ed), (1990). WordNet: An On-Line Lexical Database. Volume 3(4) of the International Journal of Lexicography. Oxford University Press.

Miller, G.A. (1995). WordNet: An on-line lexical database. Communications of ACM, 38(11).

Murata, M., Ma, Q., Uchimoto, K., Isahara, H. (1999). An example–based approach to Japanese-to-English translation of tense, aspect, and modality. In Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (pp. 66--76). University College, Chester, England.

Nakhimovsky, A. 1987. Temporal reasoning in natural language understanding: The temporal structure of the narrative. Proceedings of the Third Conference of the European Chapter of The Association for Computational Linguistics (pp. 262--269). Copenhagen: University of Copenhagen.

Palmer, M. & Wu, Z. (1995). Verb Semantics for English-Chinese Translation. Machine Translation. 10(1-2), 59--92.

Paprotté, W. (1988). A discourse perspective on tense and aspect in Standard Modern Greek and English. In B. Rudzka-Ostyn (Ed.), Topics in cognitive linguistics (pp. 447--505). Amsterdam/Philadelphia: John Benjamins.

Pustejovsky, J. (1991). The Syntax of Event Structure. Cognition, 41, 47--81.

Reichenbach, H. (1947). Elements of symbolic logic. Berkeley: University of California Press.

Resnik, P. (1992). WordNet and distributional analysis: A class-based approach to lexical discovery. In Workshop on Statistically-Based Natural-Language-Processing Techniques, San José.

Sandford Pedersen, B. (1999). Systematic Verb Polysemy in MT: A Study of Danish Motion Verbs with Comparisons to Spanish. Machine Translation 14, 39--86.

Santos, D. (1992). Tense and aspect calculus. In Proceedings of Coling 92 (pp. 1132--1236), Nantes.

Skorupka, S., Auderska, H., Łempicka, Z. (1969). (Eds.), Mały Słownik Języka Polskiego. Państwowe Wydawnictwo Naukowe, Warszawa.

Stanisławski, J. (1986). The Great English-Polish Dictionary. Wiedza Powszechna, Warszawa .

Sweetser, E.E. (1987). Metaphorical Models of Thought and Speech: a comparison of historical directions and metaphorical mappings in the two domains. In Proceedings of the 13th Annual Meeting of the Berkeley Linguistics Society.

Talmy, L. (1988). The Relation of Grammar to Cognition. In Rudzka-Ostyn, B. (Ed.), Topics in Cognitive Linguistics (pp.165--205). Amsterdam/Philadelphia: John Benjamins.

Vossen, P. (1998). EuroWordNet. A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.