Domain-related Annotation of Polish Spoken Dialogue Corpus LUNA.PL

Agnieszka Mykowiecka, Katarzyna Głowińska, Joanna Rabiega-Wiśniewska

Institute of Computer Science, Polish Academy of Sciences
J. K. Ordona 21, 01-237 Warsaw, Poland
agn@ipipan.waw.pl, k.glowinska@gmail.com, jrw@cereza.pl

Abstract

In this paper we present a corpus of Polish spoken dialogues annotated on several levels, from transcription of dialogues and their morphosyntactic analysis, to semantic annotation. The corpus is one of the results of LUNA project. The description is concentrated on the semantic annotation on the levels of concepts (attribute-value) and predicates (frame sets).

1. Introduction

In spite of the large interest in semantic role labelling (e.g. (Màrquez et al., 2008)), semantically annotated corpora are still not very frequent. The most well known resource is PropBank (Palmer et al., 2005) – a corpus of text annotated with information about basic semantic roles in which predicate-argument relations were added to the syntactic trees of the Penn Treebank. Another example of a corpus annotated with verb frames is FATE (Burchard and Pennacchiott, 2008) consisting of 800 entailment pairs from the RTE-2 (Recognizing Textual Entailment) Challenge test set, annotated with frame and semantic role labels. Another group of resources contain data annotated with specially designed domain specific labels. For example, MEDIA corpus is a domain specific resource containing hotel reservation dialogues (Bonneau-Maynard et al., 2005) (this corpus was used by French LUNA project members) while GENIA corpus (`http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`, (Kim et al., 2008)) contains about 2000 Medline abstracts annotated by term names (biological physical entities and some other important terms included in a specially designed ontology, (Kim et al., 2006)).

For a less explored language like Polish such data does not yet exist. The resource we present in this paper is the first semantically annotated corpus of Polish spontaneous speech. The LUNA.PL corpus of dialogues was collected at the Warsaw Transport Authority call centre – the telephone service which provides information on tram and bus connections, schedules, routes, fares etc. The resource was constructed in order to provide data needed for building a dialogue system which would replace call centre staff in answering a subset of customers queries. The corpus was constructed using a combination of rule based programmes and computer-aided manual work carried out in the following steps:

- data recording and selection,
- manual transcription,
- automatic annotation of speakers' turns,
- automatic morphological analysis followed by manual disambiguation,
- automatic annotation of syntactic chunks,
- automatic annotation at the level of domain attributes,
- manual verification of the attribute-value annotation,
- automatic annotation at the predicate level,
- manual verification of predicate annotation,
- final manual verification on all annotation levels.

The corpus consists of 500 directories. Each of them contains seven files: one with the recorded audio signal, one with its transcription and five XML files with annotations on subsequent levels. The collection procedure took place in the spring of 2007, (Marasek and Gubrynowicz, 2008). For the purpose of the project 500 dialogues were selected, converted into texts and annotated with speech related facts using Transcriber, (Barras et al., 1998). In the transcribed speech there are markers representing foreign words, spelling, syllabification, truncation, mispronunciation, non human noises and human noises in the background, pauses, hesitations, and articulatory noises such as breath, laugh, cough, etc. Proper names and acronyms were capitalized. In general, the transcription did not include punctuation marks, with the exception of the question mark which represented rising question intonation. After transcription, the set of dialogues was annotated with morphosyntactic tags. For this task we used the already existing analyzer (Rabiega-Wiśniewska and Rudolf, 2003) and manually disambiguated the results. The most important change in morphological description process was the introduction of special "proper names" POS classes, and the enrichment of the lexicon with a large set of proper names.

Morphologically annotated texts of dialogues were segmented into elementary syntactic chunks. The main word groups annotated were: verbal phrases, numeral phrases and several subtypes of nominal phrases. For noun and verb groups the main word was indicated.

In the rest of this paper we describe the domain related semantic annotation of the corpus. It consists of two levels – concept level at which around 200 attributes and their values are annotated and predicate level at which 47 frame types are recognized. We describe the domain model accepted, and the statistics over the entire annotated set of dialogues. At the end, the procedure of verification and evaluation of the annotation is presented. One dialogue consists on average of 128 words in 26 turns. Statistics over all levels of annotation are summarized in Table 1 (the annotation schema is presented in more detail in (Rodriguez et al., 2007) and (Mykowiecka et al. 2009)).

	number of types	occurrences
turns	-	13185
word forms	5146	82977
lemmas	2759	-
chunks	20	71779
concepts	206	32783
frames	47	5346

Table 1: Corpus annotation statistics (500 dialogues)

2. Concept level annotation

2.1. Domain model

The transport domain ontology was defined on the basis of data inspection. Although transport itself is a rather popular and relatively well known area, the types of questions which are asked by call centre customers and their information needs can be best learnt from real life dialogues. The process of domain model creation was incremental. The main part of the concept set was created on the basis of an analysis of 150 dialogues which were already annotated on the morphosyntactic levels. Then, after analyzing more dialogues, the appropriate changes in the model were introduced (Mykowiecka et al., 2008). The resulting ontology consists of 3 main root classes which describe: urban public transport domain, information seeking dialogues and time descriptions. A small fourth class contains some general features important to the domain.

Although reusing an existing ontology would be very desirable, no resource appropriate for the task has been found. Unfortunately, this is typical in areas which lay on the borders of many domains – in this case public transport organization, town topology and interpersonal communication. Domain specific resources are too big for the purpose and hard to use, while general ones (like SUMO Upper Ontology, (Pease et al., 2002)) are not precise enough. Thus a new dedicated ontology was built. It is data and application driven, i.e. concepts were introduced on the basis of real phrases and their possible role in typical information seeking calls was taken into account. The most important decisions that had to be made during the building of the annotation schema, concerned the level of detail represented and concept limits.

Level of detail

The defined ontology contains about two hundred concepts. It consists of the topology of classes and properties of class instances. As in LUNA project the annotation schema was flat, i.e. only one attribute-value pair can be assigned to any word sequence, we introduced concepts which represent both the objects themselves and their role within the dialogue. Thus, concepts included in the ontology were not limited to simple ones, but some complex concepts were also defined. For example, the *Transport* part of the ontology contains concepts representing places in town and locations in relation to these places. The concepts defined in the PLACE hierarchy are assigned only to those phrases representing places for which their specific role in the context cannot be given. In the case where places are a location of something, they are marked as LOCATION_xxx (xxx stands for all subvariants), when they are the beginning or ending of some trip or route, they are represented as SOURCE_xxx or GOAL_xxx. In the ontology, the property *isSource* which characterizes objects of type TRIP can take objects of type LOCATION as its value (to keep our ontology close to the concept set, we preserved SOURCE_xxx types in the ontology).

Annotation scope

One of the main problems which we encountered at the concept annotation stage, was the selection of concept limits, i.e. deciding which words should be labelled with a particular concept name. Very often it seems appropriate for a concept to be attributed to the entire phrase expressing the idea. In our initial solution many concepts, especially questions, were assigned to several words, sometimes entire utterances. This type of annotation turned out to be incompatible with frame level annotation, so we divided concepts that included verbs into smaller parts.

The most important (frequent) classes of the ontology are enumerated in Table 2. Figure 1 presents an annotation example as it is included within the corpus.

2.2. Annotation process

Labels with concept names were assigned automatically through the use of a specially designed rule-based program, (Mykowiecka et al., 2008). Restrictions included within the rules helped to assign concepts to phrases unambiguously. For example, morphological information was used to differentiate two usages of the phrase like *na Banacha* (*Banacha* could be either genitive or locative form in this context) as meaning either 'on Banacha (locative) street' (LOCATION_STR) or 'to Banacha (genitive) street' (GOAL_STR). The sequence *około siedemnastej* 'around 5 pm' was annotated as AROUND_HOUR – the alternative meaning 'around 17th minute' is highly unlikely.

Cases which were not disambiguated properly by rules were corrected manually. For example, numbers between 1 and 36 could be recognized as: tram number (TRAM), part of a tram/bus number (NUM_BUS_PART, NUM_TRAM_PART), stop number[1] (STOP_NUMB), street number (STREET_NUMB), minutes in time description (AT_HOUR_MINPART) as well as duration of the ride on public transport (RIDE_DURATION) and other time span (TIME_SPAN). Another example of regular polysemy are street/stop names. As most of the stop names come from street names, concepts assigned automatically are not always appropriate. These became the subjects of manual correction.

Analysing spoken language we encountered many typical irregularities, e.g.:

- **change of order**, e.g. *trzynasta godzina*, *godzina trzynasta* '1 pm'. For this and similar phrases separate rules reflecting all possible orders were defined. In many other cases the problem was made manageable through the annotation of shorter text fragments.

[1] Stop names consist of proper names (e.g. street name, building name) and the numbers 01, 02 etc.

- **inflectional errors**, e.g. in *przy Metro Politechnika* 'at/near Politechnika metro station' the noun phrase is in nominative instead of locative (should be: *przy Metrze Politechnika*). As this turned out to be a frequent mistake we did not put restrictions on case in prepositional phrases, which could not be ambiguous. So this phrase is recognized by a rule which requires the preposition *przy* followed by a stop name (including metro station names) regardless of its case.

- **elliptical phrases**, e.g. *na Krakowskim*[2] instead of *na Krakowskim Przedmieściu* or *siedem po* ('seven past' unknown hour).

  For phrases which cannot be disambiguated without knowing textual or situational context and unfinished phrases, we introduced special concepts, e.g. for *dalej* (that means either 'futher, long distance away' or 'then, afterwards') – LOC_TIME_REL. If the meaning of an elliptical phrase was unambiguous, we introduced the appropriate concept. For unfinished phrases, we introduced special concepts, e.g. *siedem po* is annotated as AT_HOUR_MINPART instead of AT_HOUR as there is no information about the hour itself.

- **interjections**, e.g. *przystanek w stronę tego Żoliborza* 'bus stop to – you know – Żoliborz', *około godziny, nie wiem, gdzieś ósmej rano* 'at around – I don't know – eight am'.

  In such situations concepts were recognized only if we created rules for each case. It was also possible to put optional elements into the rules but with great caution. Otherwise we would obtain concepts with some elements that do not belong to the very phrase.

- **unfinished utterances**, which may appear when one speaker breaks off in the middle of a sentence or is interrupted by another speaker.

  For such cases we introduced special concepts (with PART in concept name) to describe these pieces of information: NUM_BUS_PART, NUM_TRAM_PART, AT_HOUR_PART and AT_HOUR_MINPART.

- domain related phrases which appear in a context not represented within the domain model (it would probably be better to not annotate them at all), e.g. *Czy są* [ACTION *Be*] *na tych słupkach te rozkłady* [TIMETABLE *General*] *nowe?* 'Are there new timetables on those poles?'.

The general distribution of concepts in the corpus is given in Table 3 while Table 4 contains numbers of concepts with different numbers of values.

[2] *Krakowskie Przedmieście* is a name of the street, often used in abridged form *Krakowskie*. Because in Warsaw there exists a similar street name *Krakowska*, that has the same lemma in our dictionary of proper names but differs in gender, the appropriate rule had to depend not only on lemma but also on the gender of a form (or alternatively had to list all case forms).

vis-a-vis Hotelu Polonia jest autobus 118 w kierunku
vis-as Hotel Polonia there is bus 118 in the direction
tak / yes
Alei Niepodległości / Niepodległości Alley
i on również dojeżdża bezpośrednio do ulicy Spartańskiej
and it also goes directly to Spartańska street
<concept id="40" span="word_100..word_102" attribute= "LOCATION_BLD" value="OPPSIDE Hotel Polonia" />
<concept id="41" span="word_103" attribute="Action" value="ToBe" />
<concept id="42" span="word_104..word_106" attribute="BUS" value="118" />
<concept id="43" span="word_109" attribute="REACTION" value="Confirmation" />
<concept id="44" span="word_110..word_111" attribute= "GOAL_DIRECTION_STR" value="Aleja Niepodległości"/>
<concept id="45" span="word_113" attribute= "MEANS_OF_TRANSPORT" value="General" />
<concept id="46" span="word_115" attribute="Action" value="Approach" />
<concept id="47" span="word_116" attribute= "CONNECTION_Q" value="Direct" />
<concept id="48" span="word_117..word_119" attribute="GOAL_STR" value="Spartańska" />

Figure 1: Attribute level annotation example

3. Predicate level annotation

3.1. Frames definitions

The next level of annotation represents the predicate–argument structure. Our description consists in linking a verb with its compliments in one frame and is based on the FrameNet proposal, (Baker et al., (1998)). This annotation is based on the previous one, i.e. frame slots can only be filled by concepts. A verb constitutes a frame set which is filled by frame elements defined by single concepts or a sequences of concepts. This makes annotation easier, but sometimes leads to data loss as in the following example: *tam do na do Carrefoura* 'there to for to Carrefour' in which two concepts LOCATION_REL (*tam*) and GOAL_BLD (*do Carrefoura*), which could be annotated as representing frame slot GOAL are separated by *'do na'*).

At the first step of predicate annotation, 84 most frequent verbs have been described with one to six (for *być* 'to be' and *mieć* 'to have') frames. Then, frame proposals were automatically generated and manually disambiguated. In the last step of annotation we added frame descriptions for less frequent verbs as well as some new frame definitions (domain related usage of 140 verbs is described).

In total, 47 frames have been defined. In the set of frames there are three well formed thematic groups. The main one is built from verbs of movement and consists of 15 frames. Two other frame groups are formed from verbs describing places (6 frames) and verbs related to recognition/statement (5 frames). Frames which are represented in dialogues by the highest number of verb lexemes are: MOTION built from 17 verbs, SELF_MOTION represented by 16 verbs, and ARRIVING – 15 verbs.

Most of the frames are the same as defined in the English FrameNet, (Fillmore et al., 2003). In many frames new

class path	occurrences
TRANSPORT	
MEANS_OF_TRANSP	519
LOCATION	
> LOCATION_ABS	1382
> LOCATION_REL	1263
PLACE	1579
TRANSP_LINE_FEAT> DEPARTURE	192
TRIP_FEATURE	
> GOAL	1777
> GOAL_STP	276
> GOAL_STR	635
> SOURCE	1228
> SOURCE_STP	277
> SOURCE_STR	400
> PATH	377
> CONNECTION_Q	256
TRANSPORT DIALOGUE	
CONV_FORM	1657
REACTION	5046
QUESTION	
> Q_CONF (confirmation)	1040
> Q_WHAT_PLACE	84
> Q_WHAT_TIME	135
> Q_WHERE	224
TIME	
TIME_POINT	
> TIME_POINT_ABS	1144
> TIME_POINT_REL	448

Table 2: Most frequently occurring concepts

	nb	% of concepts	occurrences	% of all
>=500	12	6.03	21800	66.50
100-499	32	16.08	7301	22.27
50-99	28	14.07	1903	5.80
10-49	65	32.66	1569	4.79
0-9	62	31.16	210	0.64
Σ	199		32783	

Table 3: Concepts' frequencies

frame elements have been introduced, e.g. in the START frame, which was used for the description of the verb *zaczynać się* ('start ' in 'a street starts from') instead of the element EVENT we chose THEME. Some completely new frames were also defined, e.g. ENTITLED frame which describes fare discounts and GET frame created to present finding one's way.

For describing verb frames, we used the concepts defined in the previous level of annotation. Sometimes this resulted

concepts with # of values	concept types	concept occurrences	different values
above 100	12	6459	2264
from 51 to 100	7	8328	492
from 11 to 50	45	4958	979
up to 10	136	13038	486

Table 4: The overall distribution of concepts' values

in an unexpected context for a given verb, as in the phrase *dziecko bezpłatnie jeździ autobusami* 'a child goes by bus for free', which gives us information about fare regulation not trip description. In such cases we chose the best fitting frame to the utterance, in the presented example – ENTITLED: BENEFICIARY, PRIVILEGE, CIRCUMSTANCES. The same set of frame elements can be found in the description of verbs indicating fare discounts, e.g. *studenci mogą korzystać z 50% zniżki na bilety* 'students may use 50% discounts on tickets'. The secondary meaning of a given verb lead us also to the frame change during the last stage of annotation. For example phrases built by a verb *interesować* 'to interest' were first annotated with a frame INTEREST. The usage of that verb points out an aspect of getting a piece of information (e.g. verbs 'to ask', 'to learn', 'to find out'). Thus, we changed the frame description of the verb *interesować* to the frame TOPIC.

In Polish there are also predicates represented by two words, frequently they are constructed by an auxilliary verb *być* 'to be' or *mieć* 'to have' and a predicative noun or adjective, e.g. *mieć przystanek* ('to have a stop' meaning 'to stop'), *być czynnym* ('to be operational' meaning 'to be open') or *mieć pytanie* ('to have a question' meaning 'to ask'). Descriptions of complex predicates depended on their occurrence in text: if predicate parts are continuous, then the whole predicate is treated as the head of a predicate set, e.g. *mieć pytanie* constitutes frame TOPIC, and *być czynnym* – BEING_OPERATIONAL. In the HALT frame describing the phrase *mieć przystanek* we introduced POSITION for representing *przystanek* 'a stop', as these words were frequently separated by others.

Fillmore's frames cannot fully reflect the complexity of Polish prefixal verbs. However, from the other point of view they give the opportunity for generalizing several verbs. For example, we represented *jechać* 'to go' as RIDE_VEHICLE or SELF_MOTION, *dojechać* 'to arrive' – ARRIVING, *odjechać* 'to depart' – DEPARTING, *podjechać* 'to drive up' – MOTION, *pojechać* 'to go' – MOTION, *przejechać* 'to pass' – MOTION or TRAVERSING, *wjechać* 'to drive in' – ARRIVING, *wyjechać* 'to go away' – DEPARTING, *zjechać* 'to turn aside' as MOTION_DIRECTIONAL.

3.2. Annotation related problems

Frame annotation of transcribed speech is not easy. The most significant problems encountered were polysemy and frame selection. Moreover, we had to address problems of metonymy and discontinuities.

Regular polysemy could rarely be resolved automatically, – due to the lack of context, most cases had to be decided on manually, e.g. *jechać* can be represented by the frame RIDE_VEHICLE – if someone goes by bus or tram, or SELF_MOTION if the verb describes bus or tram motions.

The verb *być* 'to be' gives a good example of potential problems with **frame selection**. It was especially difficult to decide which frame: EXISTENCE or APPEARANCE should be chosen, as in examples A *bezpośredniego połączenia niestety nie będzie* 'unfortunately there will be no direct connection' and B *ile to jest przystanków* 'how many stops is it'. We analysed the usage of the verb *być* in different contexts, and during the annotation process we followed two

rules: an EXISTENCE frame was chosen when a predicate built a structure 'It is (not) X' (It is ENTITY), as in example A; a predicate of an APPEARANCE frame has two arguments and the structure 'X is Y' (PHENOMENON is CHARACTERIZATION), presented in example B.
Sometimes two frames could be assigned to the text but one frame can represent more information than another, e.g. for *czy eskaemką z Wesołej do Warszawy emeryci jeżdżą za darmo* 'do seniors go by (eskaemka) train from Wesoła to Warszawa without paying?' the frame ENTITLED could be used, but more elements would be represented if the frame RIDE_VEHICLE was chosen.

Metonymy is present in the example *23:12 jest tylko do przystanku Młynarska* '23:12 is only to Młynarska stop'. The time description refers to the bus which departs at that moment; but as it was not annotated as such, it cannot be used as a filler of the THEME slot.

Discontinuous phrases required some changes in description or leaving part of information unannotated, e.g. the phrase *przejazdy ma bezpłatne* 'rides have for free' is annotated as the ENTITLED frame. If the nominal phrase were continuous, *przejazdy bezpłatne*, it would be annotated as a PRIVILEGE slot. However, in the cited order the word *przejazdy* was not annotated at all.

```
<Set id="5" subset="1" span="word_103" frame=
  "EXISTENCE" head="być">
 <Frame fe="1" span="word_104..word_106"
  concept_id="42" name="ENTITY" />
</Set>
<Set id="6" subset="1" span="word_115" frame=
  "ARRIVING" head="dojeżdżać">
 <Frame fe="1" span="word_113" concept_id="45"
  name="THEME" />
 <Frame fe="2" span="word_116" concept_id="47"
  name="MANNER" />
 <Frame fe="3" span="word_117..word_119"
  concept_id="48" name="GOAL" />
</Set>
```

Figure 2: Predicate level annotation example

Figure 2 shows the frame level annotation of the same fragment which was presented in Figure 1. In Table 5 a list of the most frequent frame types is included.

4. Verification and Evaluation

To achieve high quality corpus annotations all labels were checked by a linguist who did not take part in the previous stages of the project. For this purpose, a specially designed program to visualize all annotations and enable corrections, was used, (Marcińczuk, 2008). In Table 4. we show the number of changes which were made at this stage (AER is the ratio of the sum of deleted, inserted and confused concepts w.r.t. a correct number of concepts/frames).
An evaluation of the corpus was done on a small sample of 10 dialogues (1667 words annotated with 677 concepts and 102 frames) which were checked by two linguists. The results of this evaluation are shown in Table 7.

frame name	nb of occurrences
APPEARANCE	201
ARRIVING	772
BEING_LOCATED	506
BOARD_VEHICLE	96
CAUSE_CHANGE	68
CHANGE_DIRECTION	102
DEPARTING	168
ENTITLED	211
ESCAPING	119
EXISTENCE	871
GET	69
HALT	229
MOTION	201
POSSESION	53
REFFERING_BY_NAME	59
RIDE_VEHICLE	199
SELF_MOTION	721
STATEMENT	51
TOPIC	181

Table 5: Frames with more than 50 occurrences

	correct conc.	correct values	added	not rec.	subs.	AER
CONCEPTS	32513	30422	342	976	428	7.9
Action	7093	6143	112	594	67	16.9
REACTION	5052	4925	78	71	153	6.2
STOP_DESC	969	914	10	10	8	5.8
BUS	1631	1610	1	1	8	1.8
FRAMES	4665	-	172	35	156	7.8

Table 6: Changes made in the final verification stage

5. Summary

The LUNA.PL corpus is the first semantically annotated corpus of Polish spontaneous speech data. The data will be available for research purposes and distributed together with a book containing a detailed description (Marciniak, 2010). The corpus is meant to be used in various experiments concerning speech understanding and dialogue systems construction. The first version of the resource was already used in two experiments of automatic semantic labeling using CRF (Lehnen et al., 2009), (Mykowiecka and Waszczuk, 2009). On the transliterated speech, we achieved an F-measure value of 0.85 for the concept names.

Acknowledgements This work was supported by LUNA – STREP project in the EU's 6th Framework Programme (IST 033549) .

	nb in corpus	kappa coefficient		
	C	A1/C	A2/C	A1/A2
concept labels	808	0.96	0.96	1
frames' slots	440	0.95	0.92	0.90

Table 7: Semantic annotation evaluation

6. References

Baker, C. F. , Ch. J. Fillmore, and J. B. Lowe. (1998) The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*.

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman. (1998) Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech. In *First International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376.

Bonneau-Maynard, H., S. Rosset, Ch. Ayache, A. Kuhn, D. Mostefa, and the MEDIA consortium. (2005) Semantic annotation of the MEDIA corpus for spoken dialog. In *ISCA Interspeech*, volume ISCA Interspeech, Lisbon.

Burchard A. and M. Pennacchiott. (2008) FATE: a framenet-annotated corpus for textual entailment. In *Proceedings of LREC 2008, Marrakech, Marocco*.

Fillmore, Ch. R., Ch. J. Johnson, and M. R. L. Petruck. (2003) Background to Framenet. *International Journal of Lexicography*, 16.3:235–250.

Kim, J., T. Ohta, Y. Teteisi, and J. Tsujii. (2006) GENIA ontology. Technical report, Tsujiilab, University of Tokyo.

Kim, J., T. Ohta, and J. Tsujii. (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10.

Lehnen, P., S. Hahn, H. Ney, and A. Mykowiecka. (2009) Large scale Polish SLU. In *INTERSPEECH 2009, Brighton*. ISCA.

Marasek K. and R. Gubrynowicz. (2008) Design and data collection for spoken Polish dialogs database. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Marcińczuk, M. (2008) Manufakturzysta 2.0. Dokumentacja użytkownika.

Marciniak, M. editor. (2010) *Anotowany korpus dialogów telefonicznych*. EXIT.

Màrquez, L., X. Carreras, K. C. Litkowski, and S. Stevenson. (2008) Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Mykowiecka, A. and J. Waszczuk. (2009) Semantic annotation of city transportation information dialogues using CRF method. In *Text, Speech and Dialogue 12th International Conference, TSD 2009, Pilsen, Czech Republic, LNAI 5729*, pages 411–419. Springer.

Mykowiecka, A., M. Marciniak, and K. Głowińska. (2008) Automatic semantic annotation of Polish dialogue corpus. In *Text, Speech and Dialogue 11th International Conference, TSD 2008, Brno, Czech Republic, LNAI 5246*, pages 625–632. Springer.

Mykowiecka, A., K. Marasek, M. Marciniak, J. Rabiega-Wiśniewska, and R. Gubrynowicz. (2009) Annotated corpus of Polish spoken dialogues. In *Human Language Technology. Challenges of the Information Society. Third Language and Technology Conference, LTC 2007, Poznan, Poland, October 5-7, 2007, Revised Selected Papers, LNCS 5603*, pages 50–62. Springer.

Palmer, M., D. Gildea, and P. Kingsbury. (2005) The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*.

Pease, A., I. Niles, and J. Li. (2002) The suggested upper merged ontology: A large ontology for thesemantic web and its applications. *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*.

Rabiega-Wiśniewska J. and M. Rudolf. (2003) Towards a Bi-Modular Automatic Analyzer of Large Polish Corpora. In R. Kosta, J. Błaszczak, J. Frasek, L. Geist, and M. Żygis, editors, *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages – FDSL IV, held at Potsdam University, November 28-30th, 2001*, pages 363–372.

Rodriguez, K., G. Riccardi, M. Poesio, F. Bechet, G. Damnati, K. Marasek, R. Gubrynowicz, A. Mykowiecka, J. Wisniewska, and C. Popovici. (2007) *Specifications of the annotation protocol for the data*. LUNA project deliverable D1.3. http://www.ist-luna.eu/.