Ontology-based Generation of Event Extraction Templates and Frames

Jolanta Cybulka, Jakub Dutkiewicz and Michał Żętkowski

Poznań University of Technology
3B Piotrowo Street, Poznań, Poland
{jolanta.cybulka,jakub.dutkiewicz}@put.poznan.pl, michal.zetkowski@gmail.com

Abstract

We consider a problem of semantics-driven method of events extraction from Polish free-texts. The extraction process is governed by templates generated on the basis of a rich enough ontology, which serves as a specification of user's knowledge regarding a domain of interest. We focus on templates and frames generation method and tool but the whole architecture of the event extractor is also mentioned.

Keywords: ontology-based generation of extraction templates and frames, semantics-driven events extraction, thematic roles and their syntactic approximation

1. Events extraction from free texts

There is no need to convince anyone who uses the rich free-text repositories of Internet that the automation of knowledge acquisition from them is highly desirable. The enormous amount of factual information hidden in texts should be precisely revealed in order to fit the different user's information needs (for example Machine Translation, Question Answering, Text Summarization etc. (Piskorski and Yangaber, 2013)). Such a task is known as Information Extraction (IE) and, in general, relies on obtaining instances of predefined types of entities (including relationships and events) of some domain of interest from free-texts. Apart entities themselves, their arguments or characteristics are also extracted. The obtained target instances are *structured* (in some knowledge representation format) as the opposite to *unstructured* source natural language texts. The presented specification of IE process implies the necessity to consider at least two issues: on the one hand it is the representation of domain's semantics (i.e. the predefined types of entities and their role in knowledge representation) and on the other hand – the identification of possible syntactic representations of entities in unstructured texts. Both issues have their specificities but the second is particularly difficult to address, especially if we deal with highly inflected languages such as Polish. Also, the entity description may not appear in only one sentence but in many sentences and this brings about further difficulties in natural language analysis.

In the paper we concentrate on a specific form of extraction, namely the Event Extraction (EE), in which we detect "what happened" and "what were the circumstances" of it. We consider this kind of extraction as domain-dependent – and specify domain's semantics via a c.DnSPL ontology (see section 2) that defines the user's (text reader's) predefined knowledge concerning events (thus we do not deal with *open information extraction*). This (conceptual) knowledge may have many syntactical representations in Polish, to cope with it we propose a method of generation (from the ontology) of information extraction templates that contain possible syntactic representations of events' semantics (see section 3). To do so a mapping between semantics and syntax should be established – in our case thematic roles used as parameters of *perdurants* (entities that "happen") are identified with their syntactically approximated equivalents proposed in (Jaworski and Przepiórkowski, 2014). In section 3.2 we present the method of templates generation supported by the tool. The generation of templates is accompanied by the generation of knowledge frames. The generated templates and knowledge frames were used in an extraction process briefly described in section 4. This is the domain-independent linguistic part of the extractor. In section 5 we conclude our paper.

2. Specifying reader's knowledge by means of an ontology

The domain-dependent part of our events extractor is based on semantics given via an ontology, which is based on the foundational ontological pattern of *constructive descriptions and situations* of (Gangemi et al., 2007). This pattern delivers a general view on reality and can be applied to conceptualize any domain. We adopted it to create a bilingual (Polish-English), layered and capsular c.DnSPL ontology (Cybulka, 2015). A capsule specializes (in terms of subsumption) the foundational pattern representing a "situation" that forms a conceptual equivalent of some domain. The capsule skeleton is universal and it has 9 components, from which we describe here only those that are necessary to understand our ideas. For example, let us consider a domain of terrorist incidents as it is understood in MUC-4 (Proceedings MUC-4, 1992) knowledge frame. To represent it we created a capsule (Fig. 1) named *c.DnSPL relation of the situation of a terrorist incident according to MUC-3 and MUC-4* and two typologies of ontological entities: *Typology of ground entities constituting a situation of a terrorist incident* and their equivalent *Typology of Concepts classifying entities that constitute a situation of a terrorist incident*. The latter are results of perceiving of the former by some *Agent that perceives a situation of a terrorist incident*. Among the *Typology of Concepts classifying…* the *perdurants* called *tasks* are of special interest for they represent events. Let us consider the ontological nature of a kidnapping event, reified as a *Task of kidnapping (MUC-4 slot4 INCIDENT:TYPE)*.

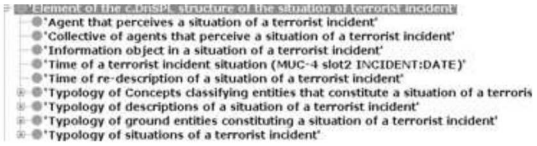

Fig. 1: The exemplary capsule of terrorist incidents as seen in Protégé editor

To serve as a semantic basis for extracting such events from free-texts the *perdurant* should have parameters such as: a perpetrator, victim, location in time and place, and optionally a beneficiary, source and target places, manner, result etc. These parameters are modelled by using *thematic roles*, respectively a role of an: *agent*, *patient-object*, *location*, *patient-beneficiary*, *ablative location*, *allative location*, *manner*, *result* (see Fig.2).

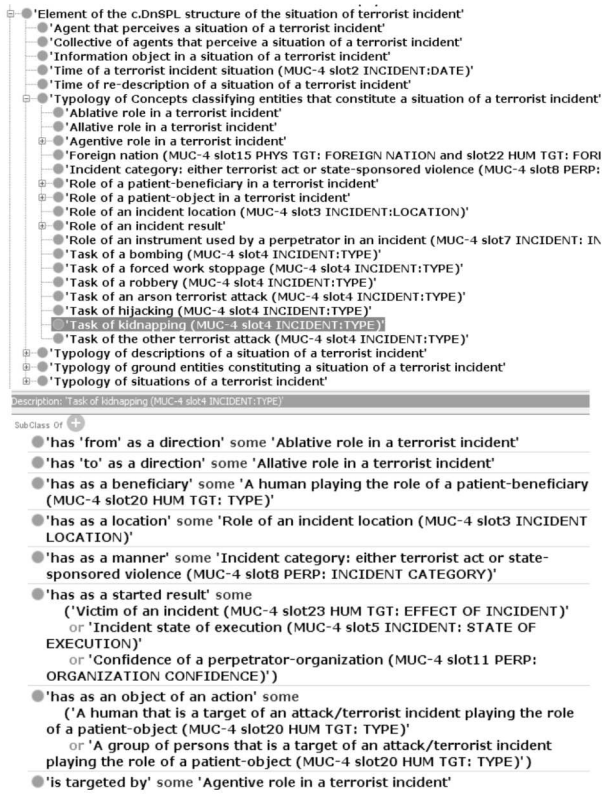

Fig. 2: The event of *kidnapping* (up) and its description (down) as seen in Protégé editor

The considered *thematic roles* are concepts used to classify the real domain entities represented as *Typology of ground entities* …. For example, a civilian may play the role of a victim-patient-object (Fig. 3.) while a terrorist organization may play the role of a perpetrator-agent. All these concepts should be carefully lexicalized, for instance with the use of a Polish WordNet.

Having such a knowledge scheme in mind one may expect that the considered ontological capsule specifies the semantics of the italicized part of a NJKP[1] sentence "*Krystynę Starczewską porwano z ulicy do pałacu Mostowskich*, pokazano nakaz ekstradycji narzeczonego córki i zaproponowano targ – współpraca za pozostawienie Nama w spokoju.", which means "*Krystyna Starczewska was kidnapped from the street to Mostowski Palace*, …". The verb *porwać/kidnap* is here in impersonal form, so there is no an instance classified as *Agentive role in a terrorist incident* but the proper name of a woman *Krystyna Starczewska* (in accusative; it may be an example of a concept *Human* that acts for an instance of *Civilian*, Fig.3) suggests she plays the *Role of a patient-object in a terrorist incident*, which is specialized by *A human that is a target of an attack ... (MUC-4 slot20 HUM TGT: TYPE)*, Fig. 2.

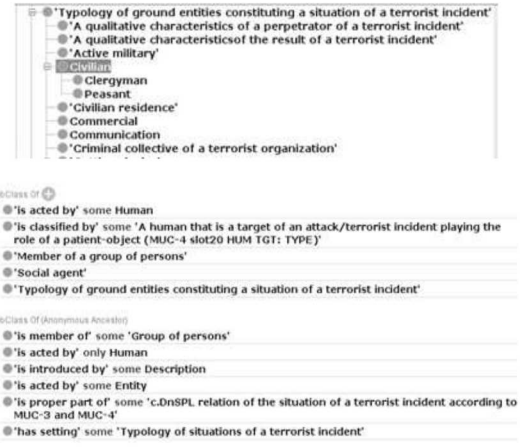

Fig. 3: The ground entity of *civilian* (up) and its description (down) as seen in Protégé editor

Also, the prepositional phrases *z ulicy*/*from the street* and *do pałacu Mostowskich*/*to Mostowski Palace* are expressions of ablative and allative locations.

Consider then another NJKP sentence "Jeden z nich w rozmowie lekko odwrócił się od stołu, a wtedy *orzechówka porwała mu z talerzyka ogromny kawał kiełbasy* i uciekła." where the italicized part says: "*…the spotted nutcracker grabbed from his plate a big piece of sausage* …". In Polish, *porwać* is polysemous and also means *to grab*. It seems obvious that this fact is semantically inappropriate and should not be extracted in Polish due to unfitness of verb's parameters-roles: *spotted nutcracker* cannot be classified as *Agentive role in a terrorist incident* (i.e. person, organization etc.), sausage is not considered to be a kidnapped victim (i.e. person, people etc.), also a *plate* cannot be an ablative location for people. Thus the semantics specified in the ontology allows us to filter the inadequate facts. But, as it was said before, the semantics may be "dressed" in many syntactic forms. The question arises, how to provide them? To specify the possible syntactic expressions of the underlined semantics we use role approximations proposed in (Jaworski and Przepiórkowski, 2014). With the help of them we built equivalents between thematic roles and syntactic *theta-roles* (parameters of verbs in valence structures). In the first column of Table 1 we list 8 roles used in c.DnSPL ontology, to which we assign theta-roles assuming that the verb is in active voice (we have also analogous equivalents in cases of passive voice, impersonal form and a gerund). Theta-roles are

[1] NJKP – national corpus for Polish, http://nkjp.pl/index.php?page=0&lang=1.

represented using a notation from Walenty[2], the Polish valence dictionary (we assume that the used abbreviations, as *subj*, *obj*, *np* and *prepnp* are widely known; the same concerns grammatical cases). The syntactic expression of the agentive role is simplified for we assume the noun phrase to be in nominative. The same concerns patient-object – it should be in accusative.

Thematic role	Theta-role
Agentive	subj{np(str/nom)}
Patient-beneficiary	obj{np(dat)}, {prepnp(dla,gen)}, {prepnp(wobec,gen)}, {prepnp(przeciw,dat)}
Patient-object	obj{np(str/acc)}
Instrumental	{np(inst)}
Allative	{prepnp(do,gen)}, {prepnp(ku,dat)}, {prepnp(między,acc)}, {prepnp(na,acc)}, {prepnp(nad,acc)}, {prepnp(pod,acc)}, {prepnp(po,acc)}, {prepnp(pomiędzy,acc)}, {prepnp(ponad,acc)}, {prepnp(poza,acc)}, {prepnp(przed,acc)}, {prepnp(w,acc)}, {prepnp(za,acc)}
Ablative	{prepnp(dzięki,dat)}, {prepnp(od,gen)}, {prepnp(spod,gen)}, {prepnp(spośród,gen)}, {prepnp(wskutek,gen)}, {prepnp(z,gen)}, {prepnp(zza,gen)}
Locative	{prepnp(koło,gen)}, {prepnp(poniżej,gen)}, {prepnp(wokół,gen)}, {prepnp(wsród,gen)}, {prepnp(u,gen)}, {prepnp(między,inst)}, {prepnp(nad,inst)}, {prepnp(pod,inst)}, {prepnp(pomiędzy,inst)}, {prepnp(ponad,inst)}, {prepnp(przed,inst)}, {prepnp(za,inst)} {prepnp(na,loc)}, {prepnp(po,loc)}, {prepnp(poza,loc)}, {prepnp(przy,loc)}, {prepnp(w,loc)},
Perlative	{prepnp(bez,gen)}, {prepnp(poprzez,acc)}, {prepnp(przez,acc)}, {prepnp(z,inst)}

Table 1. Thematic roles and their syntactic equivalents

[2] Walenty, http://zil.ipipan.waw.pl/Walenty.

3. Ontology-based generation of event extraction templates

Having the mapping between thematic and theta-roles we are ready to give the method of ontology-driven generation of events extraction templates that form the input to the extractor. In section 3.1 we specify the language in which the templates are expressed. In section 3.2 the principles of templates generation are described and the generation tool is presented.

3.1. Language of extraction templates

The input to the event extractor consists of two parts: a *knowledge frame* and a *set of extraction templates*. The main part of a knowledge frame is a set of pairs: a *slot* and a corresponding *list of semantic types* – concepts taken from the ontology. Thus, the frame abbreviates the semantics of an event. Let us look at the frame representing a kidnapping (Fig. 4). The agentive role (*RolaAgentywnaWPrzebieguIncydentu*) forms the first slot and the semantic constraints imposed on role's syntactic realizations say that they should be: a terrorist organization (*OrganizacjaTerrorystyczna*) or a person (*Osoba*). The other slots correspond to the following roles: patient-object, patient-beneficiary, ablative, allative and locative.

```
Typ: porwać
DomyślnyCzas: true
DomyślneMiejsce: true
Slot:
UriRelacji: RolaAgentywnaWPrzebieguIncydentu;
DozwoloneTypySemantyczne: OrganizacjaTerrorystyczna,Osoba;
Slot:
UriRelacji: RolaPacjensa-ObiektuWPrzebieguIncydentu;
DozwoloneTypySemantyczne: BylyWojskowy, Dyplomata,
PrzedstawicielWymiaruSprawiedliwosci, BylyPrzedstawicielWladzy,
PrzedstawicielWladzy, PrzedstawicielAparatuPrzymusu, Straznik,
Cywil, GrupaOsob, Polityk, Wojskowy;
Slot:
UriRelacji: CzlowiekWRoliPacjensa-BeneficjentaWPrzebieguIncydentu;
DozwoloneTypySemantyczne: Osoba;
Slot:
UriRelacji: RolaAblatywnaWPrzebieguIncydentu;
DozwoloneTypySemantyczne: Lokalizacja;
Slot:
UriRelacji: RolaAdlatywnaWPrzebieguIncydentu;
DozwoloneTypySemantyczne: Lokalizacja;
Slot:
UriRelacji: RolaMiejscaZajsciaIncydentu;
DozwoloneTypySemantyczne: Obszar-MUC4-LOCATION;
```

Fig. 4: The knowledge frame to represent the kidnapping event (with the possible verb's lexicalization *porwać*)

Every *extraction template* belongs to one of four groups, from which three are singled out according to the form of a verb (here verbs are the so-called *anchoring phrases*): we have then active (PL ACT), passive (PL PASS) and impersonal (PL IMPS) templates. The fourth type of templates is anchored by the gerund noun phrase (PL GERUND). Let us look at the kidnapping template anchored by the verb *porwać* in active voice (Fig.5). We have a template *header* and a *list of elements* specifying syntactic parameters of the anchoring verb. The header starts with the template name (*Nazwa szablonu*), an associated knowledge frame (*Rama wiedzy*), a version of a template (*Wersja szablonu*), a specification of a verbal anchor and its voice (*Kotwica* and *Typ frazy*). Analyzing the ontological description of a *Task of kidnapping*... (Fig. 2) we see that the semantics of a kidnapping *perdurant* is specified by using 8 thematic roles, from

which 6 are contained in Table 1. The roles are ontological concepts that are linked to the *perdurant* by carefully distinguished, formal foundational relations. For example, the considered event is linked to the *Agentive role in…* through the relation *is targeted by*. Following this relation the first *Element* was generated: the one connected with the *Agentive role in…* (*RolaAgentywnaWPrzebieguIncydentu*) that in free-texts is represented by the sentence subject (a noun phrase *NG* in *nominative* case (*Przypadki*) without prepositions (*Przyimki*)). The other *Elements*, similarly to the knowledge frame, correspond to patient-object, patient-beneficiary, allative, ablative and locative roles. More details concerning the method of obtaining templates is given in the next subsection.

Nazwa szablonu: porwać PL ACT
Rama wiedzy: porwać
Wersja szablonu: 1
Szablon aktywny: tak
Kotwica:
Typ frazy: VG
Strona czasownika: aktywna
Element:
NazwaSlotu: RolaAgentywnaWPrzebieguIncydentu
Przypadki: nominative
Przyimki:
Typ frazy: NG
Element:
NazwaSlotu: RolaPacjensa-ObiektuWPrzebieguIncydentu
Przypadki: accusative
Przyimki:
Typ frazy: NG
Element:
NazwaSlotu: CzlowiekWRoliPacjensa-BeneficjentaWPrzebieguIncydentu
Przypadki: dative
Przyimki:
Typ frazy: NG
Element:
NazwaSlotu: RolaAdlatywnaWPrzebieguIncydentu
Przypadki: dative
Przyimki: ku
Typ frazy: NG
Element:
NazwaSlotu: RolaAblatywnaWPrzebieguIncydentu
Przypadki: dative
Przyimki: dzięki
Typ frazy: NG
Element:
NazwaSlotu: RolaMiejscaZajsciaIncydentu
Przypadki: instrumental
Przyimki: między,nad,pod,pomiędzy,ponad,przed,za
Typ frazy: NG

Fig. 5: The exemplary template to extract the kidnapping event (with the possible verb's lexicalization *porwać*)

3.2. Template generation rules and tool

The method of generation can be characterized by the following scheme, which is later illustrated by screenshots of templates generation tool.

Input: Ontology in OWL
Output: Knowledge frame, set of extraction templates
Method:

1. Parse the ontology and check if it is a capsular c.DnSPL – if not, stop.
2. For each capsule do
   list its *perdurants* (i.e. tasks)
3. For a user chosen *perdurant* do
   a) read-in a verb in infinitive
   b) generate active voice templates
   c) generate passive voice templates
   d) generate impersonal form templates
   e) generate a gerund template
   f) generate knowledge frame

Fig. 6 shows an attempt to parse a non-capsular OWL ontology. After the capsular c.DnSPL ontology is parsed its capsules and *perdurants* in them are listed. The task of kidnapping is highlighted in Fig. 7. The user starts generating templates giving the lexicalization *porwać* for the kidnapping event (Fig. 8). The generation of templates relies on exploring the description of a task. The generator looks for thematic roles and if they are found in the description it generates on output their syntactic equivalents, for active voice (and impersonal form) they are given in Table 1.

Fig. 6: Non-capsular c.DnSPL ontology on input (*Brak kapsułek c.DnSPL w ontologii*)

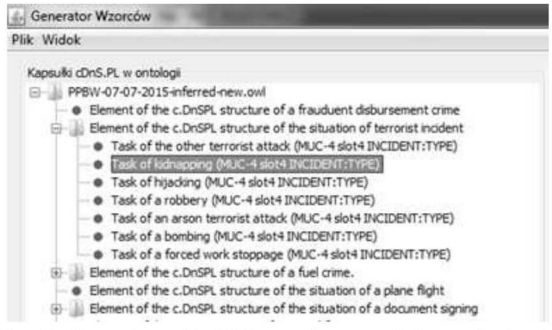

Fig. 7: Capsular c.DnSPL ontology and the tasks in the capsule of terrorist incidents

Fig. 8: Providing the generator with a verb for a chosen *perdurant*

Considering all variants of theta-roles, maximally, i.e. if all 8 roles are used in a task description, we have 162 templates for one verb form plus one GERUND template. In the description of kidnapping we have 7 roles and 163 templates in all. The left panel of the screenshot from Fig.9 shows the capsules and *perdurants* of the parsed ontology, where the task of kidnapping is highlighted. The right upper panel contains the generated templates and a knowledge frame after providing the verb and using *Generuj* button. The right lower panel shows the system's logs.

4. Event extractor architecture

The architecture of the whole EE system is sketched in Figure 10. The *Event extractor* module is a domain-independent, purely linguistic part of our extraction

system. It uses the domain-dependent *Ontology* resource which contains the lexicalized concepts.

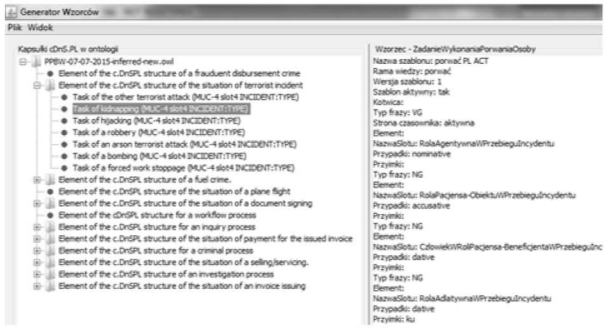

Fig. 9: The screenshot with generated templates

It uses the domain-dependent *Ontology* resource which contains the lexicalized concepts. These lexicalizations serve as a necessary, supportive tool in the extraction process. The *Event extractor* makes use of the generated templates (*Event templates database*) and the generated frames (*Event frames database*) to retrieve information from the free-text messages. Each template corresponds to one valence structure of the sentence. The messages are read through the *GUI* and then they are morphologically and syntactically processed by the *Tagger and tokenizer TaKIPI* (Piasecki, 2007) and the *Shallow Parser Spejd* (Przepiórkowski, 2008).

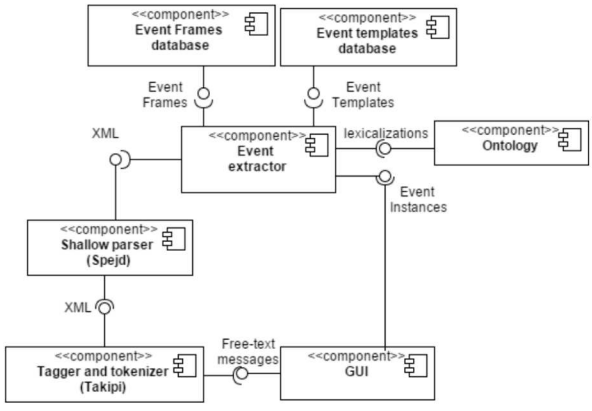

Fig. 10: Architecture of the event extraction process expressed via the UML component diagram

We ran several quality tests with the generated event templates and learned that the automatic generation process is a vast improvement to the extraction method. First of all, generated templates cover all possible valence structures of the sentences. The creation of all the usable templates is an extremely time consuming process for a human. With a large number of templates, the extractor generates a lot of results, some of which are not complete. To recognize the correct answer, we need to compare the number of extracted thematic roles between the template and the resulting extracted instance. If the numbers match, the template with the highest number of thematic roles is considered to be the correct one. That is an additional knowledge, which might be useful in the valence structure and information retrieval research. The method works for the Polish language, it provides the additional, holistic information of the valence structure of the investigated sentence, and, fortunately, it does not affect the time consumption of the extraction, since parsing and tagging subprocesses are still the bottleneck of the method. Also, the proposed method saves the preliminary time of creating event templates.

5. Final remarks

The contribution of this paper is the method of generating of event extraction templates, which is a domain-dependent process. The domain knowledge is represented as a capsule of a c.DnSPL (Cybulka J., 2015) ontology that uses thematic roles as concepts. Such roles are used in purely linguistic valence lexicons for different languages but are not used as concepts in a domain-ontology. We implemented a tool that parses an ontology and produces templates using a mapping between semantic thematic roles and syntactic theta-roles, which was inspired by (Jaworski, Przepiórkowski, 2014). The system works for Polish. We also mention a method and tool for event extraction that is described in (Cybulka, J., Dutkiewicz J., 2015). The problem of quantity evaluation of our method still stands before us – its solution requires a proper and full lexicalization of our ontology with the use of, for example, *Słowosieć* resource for Polish (http://plwordnet.pwr.wroc.pl/wordnet/).

References

Cybulka, J. (2015). *The OWL version of c.DnSPL ontology*. Retrieved from: http://users.man.poznan.pl/jolac/PPBW-22-07-2015-inferred-new.owl. Access date: 01-11-2015 (20 MB).

Cybulka, J., Dutkiewicz J. (2015). Events Extractor for Polish in a Semantics-Driven Mode. Submitted to LTC'2015, 7th Language and Technology Conference, November 27-29 2015, Poznań, Poland.

Gangemi, A., Lehmann, J., Catenacci, C. (2007). Norms and plans as unification criteria for social collectives. In: *Proc. of Dagstuhl Seminar 07122, Normative Multi-agent Systems*, vol. II, pp. 48–87, ISSN 1862-4405.

Jaworski, W., Przepiórkowski, A. (2014). Syntactic Approximation of Semantic Roles. PolTAL 2014. In: A. Przepiórkowski, M. Ogrodniczuk (Eds.): *Advances in Natural Language Processing - 9th International Conference on NLP*, Lecture Notes in Computer Science 8686, Springer, ISBN 978-3-319-10887-2, pp. 193-201.

Piasecki, M. (2007): Polish Tagger TaKIPI: Rule Based Construction and Optimisation. TASK Quarterly 11 No 1–2, 151–167.

Piskorski, J., Yangaber, R. (2013). Information Extraction: Past, Present and Future, Chapter 2 in: T. Poibeau et al. (Eds.) *Multi-source, Multilingual Information Extraction and Summarization 11* (Book series: Theory and Applications of Natural Language Processing), ISBN: 978-3-642-28568-4, Berlin Heidelberg-Springer, pp. 23-49.

Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992.

Przepiórkowski, A. (2008). Powierzchniowe przetwarzanie języka polskiego, Akademicka Oficyna Wydawnicza EXIT, Warszawa (in Polish).