COGNITIVE STUDIES | ÉTUDES COGNITIVES, 16: 119–132
Warsaw 2016

DOI: 10.11649/cs.2016.011

MICHAŁ MARCIŃCZUK, **MARCIN OLEKSY**, & **JAN WIECZOREK**

G4.19 Research Group, Wrocław University of Technology, Poland

{michal.marcinczuk,marcin.oleksy,jan.wieczorek}@pwr.edu.pl

TOWARDS RECOGNITION OF SPATIAL RELATIONS BETWEEN ENTITIES FOR POLISH

Abstract

In this paper, the problem of spatial relation recognition in Polish is examined. We present the different ways of distributing spatial information throughout a sentence by reviewing the lexical and grammatical signals of various relations between objects. We focus on the spatial usage of prepositions and their meaning, determined by the 'conceptual' schemes they constitute. We also discuss the feasibility of a comprehensive recognition of spatial relations between objects expressed in different ways by reviewing the existing tools and resources for text processing in Polish. As a result, we propose a heuristic method for the recognition of spatial relations expressed in various phrase structures called spatial expressions. We propose a definition of spatial expressions by taking into account the limitations of the available tools for the Polish language. A set of rules is used to generate candidates of spatial expressions which are later tested against a set of semantic constraints.[a]

Keywords: information extraction; semantic relations; Polish

[a]The results of our work on recognition of spatial expressions in Polish texts were partially presented in (Marcińczuk, Oleksy, & Wieczorek, 2016). In that paper we focused on a detailed analysis of errors obtained using a set of basic morphosyntactic patterns for generating spatial expression candidates — we identified and described the most common sources of errors, i.e. incorrectly recognized or unrecognized expressions. In this paper we focused mainly on the preliminary stages of spatial expression recognition. We presented an extensive review on how the spatial information can be encoded in the text, types of spatial triggers in Polish and a detailed evaluation of morphosyntactic patterns which can be used to generate spatial expression candidates.

1 Introduction

Spatial information describes the physical location of entities in a space. The location of entities can be encoded using absolute values in a coordinate system, or by relative references to other entities. The latter are called spatial relations. The relations can be expressed directly by spatial expressions (Kolomiyets, Kordjamshidi, Bethard, & Moens, 2013) or indirectly by a chain of semantic relations (LDC, 2008).

A comprehensive recognition of spatial relations between entities described in a text requires a complex chain of processing and reasoning including: a morphological analysis of the text, the recognition of object mentions, a parsing of the text, the recognition and classification of named entities, a co-reference resolution, and the recognition and interpretation of semantic relations. Due to this complexity, the feasibility and the quality of analysis depends on the availability of

certain tools and their performance. In this article we discuss how spatial information can be described in the Polish language and to what extent different representations of spatial relations can be recognized using the existing tools. We focus on expressions describing spatial relations between physical entities in which the spatial information is expressed mainly by a preposition.

2 Expressing Spatial Relations

Spatial information may be distributed throughout a sentence in many ways. This applies to both word classes and grammatical forms. As Levinson noticed, "this is true even in the focal cases, static description in the European languages, many of which encode important spatial distinctions in demonstrative adjectives, adverbs, spatial nominals, adpositions, cases and contrastive locative verbs" (Levinson, 2003).

We define a place as a category constructed by the grammar of a language, and its construction is done on the basis of the "distinct kinds of 'modes of anchoring and locating' that language(s) make available" (Bateman, Hois, Ross, & Tenbrink, 2010, p. 1037). The authors of the linguistically motivated ontology, the Generalized Upper Model, group them under the single concept Spatial Modality. As they conclude, "this is the part of the linguistic ontology that corresponds to the type of relationship being described in any linguistic spatial description, typically expressed grammatically by a spatial preposition, an adverb, an adjective, a part of the verb, or as entailed by the lexical semantics of the verb" (Bateman et al., 2010, p. 1049). The list is supplemented with determiners (especially demonstratives), as they provide information about an absolute (and perhaps relative) frame-of-reference (Levinson, 2003, p. 104–105).

There are similar ways of spatial information distribution in Polish language. We have divided them into two groups, as a spatial relation between objects may be indicated by lexical or grammatical triggers. The lexical triggers are prepositions, verbs, adverbs and adjectives. Although, as Levinson states, "there is a prevalent misleading presumption (...) that spatial notions are encoded primarily in just one word class, namely prepositions or postpositions" Levinson (2003), one of the main functions of prepositions (at least in Polish) is establishing spatial relation. This class is much more numerous than the others, since some relational nominals participate in the construction of complex prepositions like "z przodu" (in the front of). Another class consists of adverbs, which do not require a specified ground (such as "poniżej" (below)), although in some cases they take over the function of prepositions (for example "poniżej obrazu" (below the picture)). The next crucial part of the locative construction are locative verbs, which encode the layout of spatial objects (like "wisieć" (hang)). Finally, spatial information is also distributed throughout adjectives (both relational and qualitative). The grammatical triggers in Polish are grammatical case (mainly the genitive case) and verbal affixes. Below are examples of spatial triggers in Polish[1]:

1. Lexical:
   (a) **spatial prepositions**:
      - simple
        "wieś w Estonii"
        (*a village in Estonia*)
      - complex (with relational nominals)
        "usterzenie [na końcu] kadłuba"
        (*a tail [at the end] of the hull*)
   (b) **adverbs**:
      "tabelka [poniżej] obrazka"
      (*a table [below] picture*)

[1]Most of these examples come from Polish Corpus of Wrocław University of Technology (Broda, Marcińczuk, Maziarz, Radziszewski, and Wardyński, 2012b).

(c) **adjectives**:
"[pobliski] teren cmentarza"
(*[nearby] terrain of the cemetery*)

(d) **verbs**:
"[Opuściłem] Lizbonę"
(*I [left] Lisbon*)

2. Grammatical:

(a) **case**:
"ulice Warszawy"
(*the streets of Warsaw*)

(b) **part of verb** (verbal affixes):
"grzebień z piaskowca uniosły siły przyrody" (*the crest of sandstone was uplifted by natural forces*)

This paper focuses primarily on the spatial relations between objects. These can be understood as a specification of the relation between figure and ground (Talmy, 1983). Our approach is based on Langacker's specification of these concepts, which introduces the terms: trajector (an object that is in focus, a figure within a relational profile) and landmark (which provide points of reference for locating the trajector) (Langacker, 1987, p. 217–220).

Although there are several types of spatial triggers in Polish, the first step was to take into consideration mainly the spatial prepositions. A preliminary analysis of the corpus of geographical texts, which are filled with spatial descriptions, reveals that most of the spatial relations between the objects are triggered by prepositions. The corpus consists of 25 Wikipedia texts, taken mainly from the portal 'Geography'. They were manually annotated with the spatial triggers. The total number of spatial triggers in these documents was 915. The table below shows how many spatial expressions (in percentage terms) were constituted by each type of trigger.

Table 1: Spatial triggers in the Wikipedia geographic texts corpus

Trigger	**Percent**
preposition	52.13%
adverb	2.51%
adjective	14.64%
verb	21.42%
case	1.86%
affix	7.42%

Moreover, prepositions appear to be relatively independent triggers, which means that they are very often the only spatial trigger in the spatial expression. In many cases, a spatial expression is determined by more than one spatial trigger (e.g. "przechodzić przez" — verb + affix + preposition; "mieścić się w" — verb + preposition), but out of the all spatial expressions constituted by prepositions, 53.88% were constituted by the preposition alone. In contrast, the corresponding figure for spatial verbs is 6.75%, which means that the majority take a preposition (actually the locative, ablative, adlative or perlative prepositional phrase) in their valence frame.

A spatial relation between objects may be expressed in various sentence structures. The *trajector* and *landmark* may occur in a nominal (or adjective) phrase, which is an argument in a predicate-argument structure. For example, in the sentence: "I saw a car on the bridge", "a

car" is in focus, "the bridge" is a point of reference and both of them constitute the argument "the car on the bridge". But there is another possibility — the information about the *trajector* and *landmark* may be distributed among different phrases (arguments). For example, in the sentence: "the car is on the bridge" the first argument refers to the *trajector* and the second to the *landmark*. The presumption that the first argument of such sentences (subject) always refers to the *trajector* is misleading. There are some classes of verbs that occur with spatial expressions (e.g. a variety of locative prepositional phrases) indicating the location of an entity, which is the second argument (object) in a predicate-argument structure (e.g. Put Verbs; see Levin, 1993). It cannot be precluded that in some cases the entities, which may be represented by different arguments (e.g. both to the subject and to the object of the action) may have the same landmark expressed in a sentence by another argument. For example, in the sentence: "the rescue crew came across the body of a firefighter in the building" both "the rescue crew" and "the body of a firefighter" may be considered as trajectors. Hence, when the *trajector* and *landmark* are connected indirectly (via predicate in predicate-argument structure), a proper semantic analysis of the verb is needed.

3 Feasibility of Spatial Relation Recognition

Different ways of expressing spatial relations require specialized tools and resources in order to make the task feasible. Basic text processing, which includes text segmentation, morphological analysis and disambiguation, can be easily performed with any of the existing taggers for Polish, i.e., WCRFT (Radziszewski, 2013), Concraft (Waszczuk, 2012) or Pantera (Acedański, 2010). The accuracy of the taggers is satisfactory and varies between 89–91%.

The first step in the recognition of spatial relations is the identification of relevant entity mentions. The mentions can be: named entities, nominal phrases, pronouns and null verbs (verbs which do not have an explicit subject cf. Kaczmarek and Marcińczuk (2015)). The spans of entity mentions can be recognized using a shallow parser for Polish, i.e., Spejd (Przepiórkowski, 2008) with a NKJP grammar (Głowińska, 2012) or IOBBER (Radziszewski and Pawlaczek, 2012). Spejd recognizes a flat structure of nominal groups (NG) with their semantic and syntactic heads. A noun group preceded by a preposition is marked with the preposition as a prepositional nominal group (PrepNG). Every noun and pronoun creates a separate nominal group. The only exception is a sequence of nouns that is annotated as a single nominal group. IOBBER also recognizes a flat structure of nominal phrases (NP). A nominal phrase is defined as a phrase which is a subject or an object of a predicative-argument structure. This means, that some NP can contain several NGs. For example, "mężczyzna siedzący w piwnicy" (*a man sitting in the basement*) is a single NP that contains two NGs: "mężczyzna" (*a man*) and "piwnicy" (the basement) as a part of the PrepNG "w piwnicy". Spejd combined with IOBBER can be used to identify expressions with a spatial preposition within a single NP. According to Radziszewski (2012), the NKJP grammars, evaluated on the NKJP corpus, achieved 78% precision and 81% recall in the recognition of NGs, PrepNGs, NumNGs and PrepNumNGs. IOBBER, evaluated on the KPWr corpus, achieved 74% precision and 74% recall in the recognition of NPs (Radziszewski and Pawlaczek, 2012).

Neither parser recognizes nested mentions, i.e., "piwnica budynku" (*building basement*), which is recognized as a single mention. In fact, the phrase contains references to two entities: *building* and *basement*. Nested mentions can be recognized with the MentionDetector (Kopeć, 2014)). It uses a modified version of the NKJP grammar which can handle, to some extent, nested mentions. Ogrodniczuk, Głowińska, Kopeć, Savary, and Zawisławska (2015) provide two evaluations of mention detection on the PCC corpus: EXACT (exact mention matching) and HEAD (head mention matching). For exact matching, MentionDetector achieved 64% precision and 68% recall, and for head matching — 85% precision and 87% recall.

The next step is the categorization of entities into physical and non-physical. For nominal phrases, this can be done using a mapping between plWordNet (Maziarz, Piasecki, and Szpakowicz, 2012) and the SUMO ontology (Pease, Niles, and Li, 2002). The mapping contains more than

175,000 links between synsets from plWordNet and SUMO concepts. Other types of mentions (i.e., named entities, pronouns and null verbs) require additional processing. Most named entities are not present in the plWordNet, so they cannot be mapped onto SUMO through the mapping. However, they can be mapped by their categories, which can be recognized using one of the named entity recognition tools for Polish, i.e., Liner2 (Marcińczuk, Kocoń, and Janicki, 2013) or Nerf (Waszczuk, 2012). Liner2, for a coarse-grained model recognizing the top 9 categories, achieved 73% precision and 69% recall, and for a fine-grained model with 82 categories, 67% and 59%, respectively. Nevertheless, a mapping of named entity categories onto SUMO is required. Prepositions and null verbs also cannot be mapped through a wordnet as they do not contain any semantic information about the entity they refer to. They require a co-reference resolution to a nominal phrase or a named entity. This in turn, can be done with one of the tools for co-reference resolution in Polish, i.e., Bartek (Kopeć and Ogrodniczuk, 2012), Ruler (Ogrodniczuk and Kopeć, 2011) or IKAR (Broda, Burdka, and Maziarz, 2012a) (does not resolve co-reference for null verbs).

4 Recognition of Spatial Expressions

Kordjamshidi, Van Otterlo, and Moens (2011) present two approaches to the recognition of spatial expressions: pipeline and joint. The first approach consists of two steps: finding spatial indicators and finding spatial arguments (trajector and landmark). In the second approach, spatial indicators and their arguments are recognized jointly. The authors show that the pipeline approach outperforms the joint approach, as it takes advantage of external resources in the recognition of spatial indicators. They utilised data from the preposition project (TPP) employed in SemEval-2007. This dataset contains prepositions annotated with their senses, including the *spatial sense*. In the case of Polish, such a dataset does not exist. Thus, the pipeline approach based on machine learning is not feasible. The other issue is that Kordjamshidi et al. (2011) employ, in both approaches, a word sense disambiguation for English. In the case of Polish there is not such a robust tool. Only several experiments for a limited set of nouns have been conducted.

Taking the above into account, we decided to implement and evaluate a holistic two-stage approach. At the first stage we will use a set of morphosyntactic patterns to identify the candidates for spatial expressions. Then, a set of ontology-based constraints will be applied to filter out the non-spatial expressions. This way we can utilise the existing tools and resources which are available for Polish (see Section 3). The texts will be processed with a morphological tagger, a shallow parser, a dependency parser and a named entity recognizer. We will also use a wordnet for Polish, an ontology, and a mapping between the wordnet and the ontology. The mapping will be conducted by searching the semantic heads of phrases in the wordnet. Then, we will use a set of patterns to identify spatial expression candidates, i.e., triples containing a trajector, a preposition and a landmark. The procedure for discovering patterns is presented in Section 6.2. In the last step, the set of generated candidates will be tested against a set of semantic constraints. For each spatial preposition we will define a list of possible categories for the trajector and the landmark. The semantic constraints are described in details in Section 5.

5 Semantic constraints

We are aiming towards the automatic labelling of words or phrases in sentences with a set of spatial roles which take part in one or more spatial relations expressed by the sentence. The annotation scheme is based on the holistic spatial semantic theory (HSS) (Zlatev, 2003). The semantic spatial components in HSS theory are *trajector*, *landmark*, frame of reference, path, region, direction and motion (Zlatev, 2007).

Trajector and *landmark* identification is the most crucial task. Machine learning methods to extract spatial roles and their relations are based on the assertion that: "the sentence-level spatial analysis of texts characterizes spatial descriptions, such as determining the objects' spatial

properties and locations to answer 'what/who' and 'where' questions. The spatial indicator (typically a preposition) establishes the type of spatial relation, and other constituents express the participants of the spatial relation (e.g. 'entities' locations)" (Kordjamshidi et al., 2011, p. 4).

Information about the type of a spatial relation comes not only from the meaning of a preposition (spatial indicator). Lexemes referring to a localized object (*trajector*) and to an object of reference (*landmark*) also influence the identification of the relation denoted in a text. We can use the same preposition (in a formal sense, i.e., in a combination with the same grammatical case of a noun) to introduce information about spatial or non-spatial relations (e.g. time)). For example:

1. Piotr siedział przed domem.
   (*Piotr was sitting in front of the house.*)
2. Piotr siedział przed godziną w biurze.
   (Piotr was sitting in the office an hour ago.)

The semantic restrictions of *trajector* and *landmark* can be used to distinguish a specific meaning of the preposition due to a specific spatial cognitive pattern (Przybylska, 2002). Based on Przybylska's approach, we described these patterns using classes from the SUMO ontology and attempted to capture the prototypical conceptualization of these patterns. We have chosen the schemas that could be captured in the model: *trajector – spatial indicator landmark*, provided that the spatial indicator is a simple preposition. The set contained 160 cognitive schemes for spatial relations (including the specificity of the objects in the relation). We have focused on the spatial relations between static physical objects. Therefore, the set had to be reduced. We have excluded patterns based on motion verbs and those referring to the non-physical objects (like *informacja w książce* or *myśl w głowie*). Furthermore, we have added the schemas for complex spatial prepositions. The set was finally composed of 121 ontological schemas based on the semantic restrictions expressed with the SUMO classes. Two (out of 18) example schemas for "NA" are presented in the following subsections.

Table 2: Schema for "NA" #1

Preposition	na (*on*) #1
Interpretation	Object TR is outside the LM, typically in contact with external limit of LM by applying pressure with its weight.
Example	"książka leży na stole" (*a book is on the table*)
Class of TR	Artifact, ContentBearingObject, Device, Animal, Plant, Pottery, Meat, PreparedFood, Chain
Class of LM	Artifact, LandTransitway, BoardOrBlock, Boatdeck, Shipdeck, StationaryArtifact

Table 3: Schema for "NA" #2

Preposition	na (*on*) #2
Interpretation	TR is adjacent to a side surface of the LM, TR is visible and partially covers the LM.
Example	"plakaty na murze" (*posters on the wall*)
Class of TR	ContentBearingObject, Artifact
Class of LM	StationaryArtifact, Furniture, Clothing, Bag

6 Evaluation

6.1 Corpora

In the preliminary experiments we used two set of documents:

KPWr — a preliminary set containing 564 documents (Wikipedia articles and texts from blogs) from the KPWr corpus. The set contains 93,572 tokens and 1707 spatial relations (one relation for every 55 tokens). The documents were annotated only by one linguist.

WGT — a set of 50 geographical texts from Wikipedia. This type of article contains many spatial relations between objects. The set contains 17,407 tokens and 466 spatial relations (one relation for every 37 tokens). This set was annotated by two linguists independently and the inter-annotator agreement was measured by means of the Dice coefficient. The agreement was 82%.

6.2 Spatial Expression Patterns

Using the preliminary dataset (KPWr), we generated a list of the most frequent morphosyntactic patterns. We used Spejd and IOBBER to automatically recognize noun groups and noun phrases. Table 4 contains the patterns which appeared at least 10 times. `[NG#TR]` is a nominal group (NG) containing a trajector (TR) and `[PrepNG#LM]` is a nominal group with a preposition (PrepNG) containing a landmark (LM)." "`<....>`" represents a single noun phrase (NP). The most frequent pattern (P1) matches 29% of all expressions. For the remaining patterns the coverage drops sharply, as the next most frequent pattern (P2) matches only 3.6% of all expressions. The list of patterns generated for 1,707 expressions contains more than 570 items. 550 of them appear only once.

In initial set of patterns contains the the patterns listed in Table 4 excluding:

- P4, P5, P6 — all expressions matched by these patterns have the same form, "a village in Poland located in province ..., in community ..., in township ...". The patterns were too specific — they were characteristic for articles from Wikipedia.
- P18, P19 — the *trajector* and/or the *landmark* was a co-referential adjective that cannot be filtered by the semantic constraints without a co-reference resolution. In this research we do not cover the problem of co-reference resolution for *trajectors* and *landmarks*.

To increase the coverage, we added a pattern (P20) that is a generalization of less frequent patterns which contain both *trajector* and *landmark* in the same noun phrase separated by some other tokens (interpunctions, adjectives, adverbs, etc.). The additional tokens which do not create separate nominal groups were represented by "`...`" in `<NG|...|PrepNG>`.

The initial set of patterns was evaluated using the WGT corpus and the results are presented in Table 5. In the evaluation, we used the preliminary set of semantic constraints described in Section 5 to filter the candidates.

As can be seen, the distribution of true positives is very similar to the distribution of pattern frequency in the KPWr corpus.

After the evaluation of the initial set of patterns we introduced the following changes:

- patterns P2, P3, P9 and P15 were replaced by new patterns P21 and P22 — we found that the patterns match elements of the predicate-argument structures. A similar effect can be achieved using a dependency parser and patterns based on relations between words. A tree-based pattern might have better coverage than a linear pattern as it does not require the arguments and the predicate to follow each other. The analysis of less frequent patterns shows that there are many other patterns with a predicate which contain some additional tokens between the *landmark* and the *trajector*, for example:

Table 4: List of morphosyntactic patterns for spatial expressions with their frequency.

Id	Count	Pattern
P1	497	`<[NG#TR] [PrepNG#LM]>` “ranę na nosie” (*cut on a nose*) “Rafał z Ozorkowa” (*Rafał (first name) from Ozorków (city)*)
P2	62	`<[PrepNG#LM]> [Verbfin] <[NG#TR]>` “we Lwowie skończył się Letni Obóz” “w Ozorkowie odbędzie się mecz”
P3	44	`<[NG#TR]> [Verbfin] <[PrepNG#LM]>` “bazylika stoi na wzórzu” “Dickoh urodził się w Danii”
P4	38	`<[NG#TR] [PrepNG] [Ppas] [PrepNG] [orth=,] [PrepNG#LM]>` “wieś w Polsce położona w województwie lubelskim, w powiecie chełmskim”
P5	38	`<[NG#TR] [PrepNG] [Ppas] [PrepNG#LM]>` “wieś w Polsce położona w województwie lubelskim”

P6	37	`<[NG#TR] [PrepNG] [Ppas] [PrepNG] [orth=,]`
		`[PrepNG] [orth=,] [PrepNG#LM]>`
		“wieś w Polsce położona w województwie lubelskim
		w powiecie chełmskim, w gminie”

P7	32	`[NG#TR] [PrepNG#LM]` “profesor anatomii patologicznej w Warszawie”
P8	31	`<[NG#TR] [Ppas] [PrepNG#LM]>` “miasto położone na wyspie” “diod LED zamontowanych na płytce”
P9	25	`<[NG#TR]> <[PrepNG#LM]>` “auta wprost z pudełek” “zębatki w obudowie”
P10	20	`<[PrepNG#LM] [NG#TR]>` “we wsi karczmy”
P11	20	`<[NG#TR] [NG] [PrepNG#LM]>` “astronom Walter Frederick Gale 7 czerwca 1927 roku w Sydney” “LO im. Władysława Orkana w Sadownem”
P12	16	`<[NG#TR] [PrepNG] [orth=,] [PrepNG#LM]>` “gmina w Niemczech, w Bawarii”
P13	16	`<[NG#TR] [pos=prep] [NG#LM]>` “igrzyskach w Nagano”
P14	15	`<[NG#TR] [Pact] [PrepNG#LM]>` “akademii piłkarskiej znajdującej się w rodzinym mieście”
P15	14	`<[PrepNG#LM]> <[NG#TR]>` “(znajdująca się) we wsi parafia”
P16	14	`<[NG#TR] [PrepNG] [PrepNG#LM]>` “zajezdnia dla tramwajów miejskich w Łodzi”
P17	14	`<[NG#TR]> [Ppas] [Verbfin] [PrepNG#LM]` “miejscowość położona była w województwie”
P18	12	`<[pos=adj#LM]> [Verbfin] <[NG#TR]>` “w którym złamał się wahacz” “w każdym stoi czara”
P19	11	`<[pos=adj#TR]> [Verbfin] <[PrepNG#LM]>` “który odbędzie się na stadionie” “która zamieszkała w wiosce”

Table 5: Evaluation of patterns — first iteration

Id	Short pattern	Precision	Matched	TP	FP
P1	`<NG|PrepNG>`	101	62.38%	63	38
P2	`<PrepNG>|Verbfin|<NG>`	11	63.64%	7	4
P3	`<NG>|Verbfin|<PrepNG>`	15	40.00%	5	8
P7	`NG|PrepNG`	5	40.00%	2	3
P8	`<NG|Ppas|PrepNG>`	9	55.56%	5	4
P9	`<NG><PrepNG>`	0	0.00%	0	0
P10	`<PrepNG|NG>`	27	7.41%	2	25
P11	`<NG|NG|PrepNG>`	3	50.00%	3	3
P12	`<NG|PrepNG|Comma|PrepNG>`	2	50.00%	1	1
P13	`<NG|prep|NG>`	0	0.00%	0	0
P14	`<NG|Pact|prep|NG>`	1	100.00%	1	0
P15	`<PrepNG><NG>`	3	66.67%	2	1
P16	`<NG|PrepNG|PrepNG>`	11	9.09%	1	10
P20	`<NG|...|PrepNG>`	20	55.00%	11	9
	Total	195	48.72%	95	100

- `<[PrepNG#LM]> [Ppas] [Verbfin] [NG] [NG#TR]`,
- `<[NG#TR]> [Verbfin] <[PrepNG] [PrepNG#LM]>`,
- `<[NG#TR] [PrepNG] [Ppas]> [Verbfin] [PrepNG#LM]`,
- `<[NG#TR]> [Verbfin] [Ppas] [pos=adv] [NG] [PrepNG#LM]`.

Each of the above patterns appear only once in the KPWr corpus, but they have a common feature — `[NG#TR]` and `[PrepNG#LM]` are direct arguments of the `[Verbfin]` element. This construction can be easily represented in terms of dependency relations as:

`[NG#TR] -{obj}-> [Verbfin] <-{comp}- [Prep#SI] <-{comp}- [NG#LM]` (P21), where "`X -(Z)-> Y`" means a relation from `X` to `Y` of type `Z`.

Pattern P22 is an extended version of P21. It covers situations where there is a list of landmarks, for example "Germany, France and Poland."

In this example, the landmarks "Germany", "France" and "Poland" are connected to the preposition "and" with a conjunct relation and the preposition is connected to the spatial preposition with a comp relation. Thus, we take the pattern P21 and replace "`[Prep#SI] <-{comp}- [NG#LM]`" with "`[Prep#SI] <-{comp}- [base=i] <-{conjunct}- [NG#LM]`".

- Pattern P10 was removed — almost all the candidates matched by this pattern were incorrectly split prepositional nominal groups. For example, the phrase "w dzielnicy Krzyki" (Eng. *in Krzyki district*) was recognized as two phrases, i.e. "w dzielnicy" (Eng. *in district*) as `PrepNG` and "Krzyki" as `NG`.
- Patterns P1 and P20 were modified — we observed that most of the errors for these patterns were caused by'the incorrect recognition of nominal groups. The groups were split into two separate categories, for instance "koryto rzeki Warta" (Eng. *Warta river bed*) should be

recognize as a single nominal group but it was recognized as two groups, i.e. "koryto" (Eng. *bed*) and "rzeki Warta" (Eng. *Warta river*). The modification was made so that if the potential trajector was proceeded by other nominal groups within the same noun phrase, then the leftmost nominal group was selected as the trajector. For example, instead of `<NG NG NG#TR PrepNG#LM>` we take `<NG#TR NG NG PrepNG#LM>`. The modified patterns were named `<FirstNG|PrepNG>` (P1*) and `<FirstNG|...|PrepNG>` (P20*).

The modified set of patterns was evaluated using the same corpus as the initial set. The results are presented in Table 6. Despite the set having fewer patterns, it recognized more true positives (99 compared to 83) and fewer false positives (54 compared to 84). As expected, the replacement of linear patterns for predicative-argument structures (P2, P3, P9 and P15) with dependency-based patterns (P21 and P22) improved the recall.

Table 6: Evaluation of patterns — second iteration.

Id	Short pattern	Matched	Precision	TP	FP
P1*	`<FirstNG|PrepNG>`	26	76.92%	20	6
P8	`<NG|Ppas|PrepNG>`	9	55.56%	5	4
P14	`<NG|Pact|PrepNG>`	1	100.00%	1	0
P20*	`<FirstNG|...|PrepNG>`	88	62.50%	55	33
P21	`MALT`	58	51.72%	30	28
P22	`MALT_LMconj`	5	40.00%	2	3
	Total	170	61.18%	104	66

6.3 Semantic filtering of Spatial Expression Candidates

Table 7 shows the impact of semantic filtering on the performance. For the improved set of patterns, the number of false positives was reduced from 956 to 66, i.e. almost 20 times smaller. At the same time, the number of true positives was only reduced by one-third — from 155 to 104. Precision increased from 13.95% to 61.18%, and recall dropped from 33.26% to 22.32%.

Table 7: Complete evaluation

Pattern	Semantic filtering	Precision	Recall	F-measure	TP	FP	FN
Initial	No	12.88%	33.05%	18.53%	154	1042	312
	Yes	48.72%	20.39%	28.74%	95	100	371
Improved	No	13.95%	33.26%	19.66%	155	956	311
	Yes	61.18%	22.32%	32.70%	104	66	362

We analysed the false negatives, i.e. candidates discarded by the semantic constraints, and we identified the following main reasons:

- semantic constraints — the set of semantic constraints is incomplete and some possible concepts are missing.

- missing concepts — the system did not assign any concepts to the *trajector* or the *landmark*. There were two possible reasons. One is a missing mapping between a lexical unit and the wordnet. The other is an unrecognized named entity by Liner2.

Table 8: Type of errors causing false negatives

Type of errors	Count	%
Semantic constrains	30	58.82%
Missing concept	10	16.60%

6.4 Analysis of False Positives

We have carefully analysed all the false positives in order to identify the source of errors. We have identified the following types of errors:

- pre-processing — this type of error is caused by WCRFT (incorrect morphological disambiguation), Spejd (incorrect recognition of nominal groups), Malt (incorrect recognition of dependency relations between words or Liner2 (not recognized proper name or incorrect categorization) error.
- lack of WSD — different word senses involve different mappings on SUMO classes. In some cases, one sense may refer to an object, while another one may refer to a process. For example, in the phrase "pokój w Utrechcie" the word "pokój" was interpreted as *peace* and *room*.
- semantic constraints — this type of error is connected mainly to the generality of some cognitive schemes. The system found it difficult to distinguish between the *trajector* and the *landmark*, when they were regions, administrative units etc., one of which is a part of another (for example the system proposed the expression "Warszawie w dzielnicy" instead "dzielnicy w Warszawie")
- motion relations — semantic constraints do not distinguish between static and motion relations. In our experiment we focused only on static relations and motion relations were not annotated in the corpus. In fact, these errors will not be a problem in the future, when we also include motion relations.

Table 9: Type of errors causing false positives

Type of errors	Count	%
Preprocessing	16	30.77%
Lack of WSD	2	3.85%
Semantic constraints	21	40.38%
Motion relations	12	23.08%
Errors in corpus	1	1.92%

The analysis shows that 34 out of 66 false positives are not real errors. If we assume that the set of semantic constraints will be complemented, the corpus errors will be solved and that motion

relations will be accepted, then the number of true positives will increase to 138 and the number of false positives will decrease to 32. This gives a theoretical precision of 81% and recall of 30%.

7 Conclusions and Future Work

In this paper we have discussed the problem of spatial expression recognition in the Polish language. We have presented and evaluated a proof of concept for the recognition of spatial expressions in Polish using a holistic two-stage approach. We have focused on expressions containing a spatial preposition.

The preliminary results are promising in terms of precision — 61.18% in practise and nearly 81% in theory. Now, the main problem which still needs to be addressed is the recall of spatial expression candidates. The improved set of patterns without semantic filtering was able to discover only 30% of all expressions. This problem is a challenge, as the diversity of spatial expression patterns is very high — 550 out of 570 patterns generated on the KPWr corpus were unique. One way to solve this problem might be a generalisation of the less frequent patterns, but this is very time-consuming and tedious work. The other way is to better exploit the dependency parser.

The other problem which also needs to be addressed is the recognition of expressions which contain null-verbs adjectives as a *trajector* and *landmark*. To solve this problem, we need to utilise a co-reference resolution system which still needs improvement. In future work we also want to cover the problem of spatial expression normalization and representation in the region connection calculus (RCC-8).

References

Szymon Acedański. A morphosyntactic brill tagger for inflectional languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer, 2010. ISBN 978-3-642-14769-2.

John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. A linguistic ontology of space for natural language processing. *Artificial Intelligence*, 174(14):1027–1071, September 2010. ISSN 00043702. doi: 10.1016/j.artint.2010.05.008. `http://www.sciencedirect.com/science/article/pii/S0004370210000858`.

Bartosz Broda, Łukasz Burdka, and Marek Maziarz. IKAR: an improved kit for anaphora resolution for polish. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Demonstration Papers, 8-15 December 2012, Mumbai, India*, pages 25–32, 2012a. `http://aclweb.org/anthology/C/C12/C12-3004.pdf`.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. KPWr: Towards a free corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012b. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Katarzyna Głowińska. Anotacja składniowa NKJP. In Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors, *Narodowy Korpus Języka Polskiego*, pages 107–127. Wydawnictwo Naukowe PWN, Warsaw, 2012.

Adam Kaczmarek and Michał Marcińczuk. Heuristic algorihtm for zero subject detection in polish (to be published). In *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence. Springer Berlin / Heidelberg, 2015.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. Zero subject detection for Polish. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255–262, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Mateusz Kopeć and Maciej Ogrodniczuk. Creating a Coreference Resolution System for Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 192–195, Istanbul, Turkey, 2012. ELRA.

Mateusz Kopeć. Zero subject detection for Polish. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 221–225, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process*, 8(3):4:1–4:36, 2011. ISSN 1550-4875. doi: 10.1145/2050104.2050105. http://doi.acm.org/10.1145/2050104.2050105.

R.W. Langacker. *Foundations of Cognitive Grammar: Theoretical prerequisites*. Number t. 1 in Foundations of Cognitive Grammar. Stanford University Press, 1987. ISBN 9780804738514. https://books.google.com.mx/books?id=g4NCRFbZkZYC.

LDC. ACE (Automatic Content Extraction) English Annotation Guidelines for Relations. *Argument*, 2008.

B. Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993. ISBN 9780226475332. https://books.google.com.mx/books?id=6wIZWOrcBf8C.

Stephen C Levinson. *Space in language and cognition: explorations in cognitive diversity*. Language, culture and cognition ; 5. Cambridge Univ. Press, Cambridge [u.a.], 1. publ edition, 2003. ISBN 0521812623. http://katalog.suub.uni-bremen.de/DB=1/LNG=DU/CMD?ACT=SRCHA&IKT=8000&TRM=59363920*.

Michał Marcińczuk, Jan Kocoń, and Maciej Janicki (2013). Liner2 — A Customizable Framework for Proper Names Recognition for Polish. In Robert Bembenik, Łukasz Skonieczny, Henryk Rybiński, Marzena Kryszkiewicz, and Marek Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 231–253. Springer, 2013. ISBN 978-3-642-35646-9. http://dblp.uni-trier.de/db/series/sci/sci467.html#MarcinczukKJ13; http://dx.doi.org/10.1007/978-3-642-35647-6_17; http://www.bibsonomy.org/bibtex/2a8c58cd394c73d4b6aaa14fbc5a9c408/dblp.

Michał Marcińczuk, Marcin Oleksy, and Jan Wieczorek (2016). Preliminary Study on Automatic Recognition of Spatial Expressions in Polish Texts. In Sojka, Petr, Horák, Aleš, Kopeček, Ivan, and Pala, Karel (Eds.), *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno , Czech Republic, September 12-16, 2016, Proceedings*, pages 154–162. Cham: Springer International Publishing. DOI: http://dx.doi.org/10.1007/978-3-319-45510-5_18.

Marek Maziarz, Maciej Piasecki, and Stanisław Szpakowicz. Approaching plWordNet 2.0, January 2012.

Maciej Ogrodniczuk and Mateusz Kopeć. Rule-based coreference resolution module for Polish. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 191–200, Faro, Portugal, 2011.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter, 2015. ISBN 978-1-61451-835-8. http://www.degruyter.com/view/product/428667.

Adam Pease, Ian Niles, and John Li. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002.

A. Przepiórkowski. *Powierzchniowe przetwarzanie języka polskiego*. Problemy współczesnej nauki, teoria i zastosowania: Inżynieria lingwistyczna. Akademicka Oficyna Wydawnicza "Exit", 2008. ISBN 9788360434475. https://books.google.pl/books?id=V076OgAACAAJ.

R. Przybylska. *Polisemia przyimków polskich w świetle semantyki kognitywnej*. Universitas, Kraków, 2002.

Adam Radziszewski. *Metody znakowania morfosyntaktycznego i automatycznej płytkiej analizy składniowej języka polskiego*. PhD thesis, Politechnika Wrocławska, Wrocław, 2012.

Adam Radziszewski. A tiered CRF tagger for Polish. In H. Rybiński M. Kryszkiewicz M. Niezgódka R. Bembenik, Ł. Skonieczny, editor, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*, page to appear. Springer Verlag, 2013.

Adam Radziszewski and Adam Pawlaczek. Large-scale experiments with NP chunking of Polish. In *Proceedings of TSD 2012*, Brno, Czech Republic, 2012. Springer.

Leonard Talmy. *How Language Structures Space*, pages 225–282. Springer US, Boston, MA, 1983. ISBN 978-1-4615-9325-6. doi: 10.1007/978-1-4615-9325-6·11. http://dx.doi.org/10.1007/978-1-4615-9325-6_11.

Jakub Waszczuk. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of COLING 2012*, num-

ber December 2012, pages 2789–2804, 2012. http://cse.iitk.ac.in/users/cs671/2013/hw3/waszczuk-12coling_CRF-w-domainspecific-constraints-for-morpho-tagging.pdf.

Jordan Zlatev. Holistic spatial semantics of thai. In Eugene H. Casad and Gary B. Palmer, editors, *Cognitive linguistics and non-Indo-European languages*, pages 305–336. Mouton de Gruyter, Berlin; New York, 2003.

Jordan Zlatev. Spatial semantics. In Dirk Geeraerts and Hubert Cuyckens, editors, *The Oxford Handbook of Cognitive Linguistics*. Oxford University Press, 2007. ISBN 9780199738632. //www.oxfordhandbooks.com/10.1093/oxfordhb/9780199738632.001.0001/oxfordhb-9780199738632-e-13.

Acknowledgment

Work financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

The authors declare that they have no competing interests.

The authors' contribution was as follows: concept of the study, data collection, analyses and the writing: Michał Marcińczuk, Marcin Oleksy, Jan Wieczorek.

Publisher: Institute of Slavic Studies, Polish Academy of Sciences, University of Silesia & The Slavic Foundation