Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2167–2174
Marseille, 11–16 May 2020

PST 2.0 – Corpus of Polish Spatial Texts

Michał Marcińczuk, Marcin Oleksy, Jan Wieczorek
Department of Computational Intelligence
Wrocław University of Science and Technology
Wrocław, Poland
{michal.marcinczuk, marcin.oleksy, jan.wieczorek}@pwr.edu.pl

Abstract

In the paper, we focus on modeling spatial expressions in texts. We present the guidelines used to annotate the PST 2.0 (Corpus of Polish Spatial Texts) – a corpus designed for training and testing the tools for spatial expression recognition. The corpus contains a set of texts gathered from texts collected from travel blogs available under Creative Commons license. We have defined our guidelines based on three existing specifications for English (SpatialML, SpatialRole Labelling from SemEval-2013 Task 3 and ISO-Space1.4 from SpaceEval 2014). We briefly present the existing specifications and discuss what modifications have been made to adapt the guidelines to the characteristics of the Polish language. We also describe the process of data collection and manual annotation, including inter-annotator agreement calculation and corpus statistics. In the end, we present detailed statistics of the PST 2.0 corpus, which include the number of components, relations, expressions, and the most common values of spatial indicators, motion indicators, path indicators, distances, directions, and regions.

Keywords: information extraction, spatial expressions, data resources, annotated corpus, Polish

1. Introduction

Spatial information refers to a physical location of an object, which can be encoded using some absolute values in a coordinate system or by relative references to other entities in the space – spatial relations. The spatial relations can be expressed directly by spatial expressions (Kolomiyets et al., 2013) or indirectly by a chain of semantic relations (LDC, 2008). A comprehensive recognition of spatial relations between objects in a text requires a manually annotated corpus, which may be used as a training and testing data. Recognition and interpretation of spatial expressions is crucial to understand and reason about spatial relations between object in scene description (Chang et al., 2015) or dialogue systems (Williams et al., 2016; Marge and Rudnicky, 2019).

In the paper, we focus on modeling spatial expressions in texts for Polish. We present the guidelines used to annotate the PST 2.0 – corpus of Polish Spatial Texts. In Section 2, we present existing spatial annotation schemes for English (SpatialML, SpatialRole Labelling from SemEval-2013 Task 3 and ISO-Space1.4 from SpaceEval 2014). In Section 3 we discuss what modifications have been made to adapt the guidelines to the characteristics of the Polish language and we define basic componentes and relations used to model spatial expressions. Section 4 is an overview of the Polish Spatial Texts corpus, including the description of data source, manual annotation process and corpus statistics.

2. Existing Spatial Specifications

The specification language for spatial information in language evolves over the years. There are several schemes, which more or less correspond to each other. We based our approach on three of them: SpatialML, Spatial Role Labelling from SemEval-2013 Task 3, and ISO-Space 1.4 from SpaceEval 2014. The related annotated corpora first contained mostly static spatial expressions (Grubinger et al., 2006). The data set was extended in SemEval 2013 (SpRL task), and more dynamic spatial relations were annotated from the Degree Confluence Project (Jarret, 2019). The task was further extended by SpaceEval, and The SpaceBank Corpus was introduced (Pustejovsky and Yocum, 2013) and then re-annotated according to ISO-Space (Pustejovsky et al., 2015).

2.1. SpatialML

The SpatialML annotation scheme (Mani et al., 2010) consists first and foremost of locations marked by PLACE tags. Topological and orientation relations are also represented. There are two types of links (relations) — *RLINKS* (with *direction* and *distance* attributes) which relate relative locations to absolute ones and *SLINKS* which relates locations to each other while recording the type of topological relation involved. They are supplemented by SIGNALS tag, used for text spans that license a link. SpatialML focuses mainly on geography and culturally-relevant places, rather on other spatial language domains (Mani et al., 2008).

2.2. Spatial Role Labelling

Spatial Role Labelling is defined as 'the automatic labeling of words or phrases in sentences with a set of spatial roles' (Kordjamshidi et al., 2011). Spatial role set consists primarily of three elements: TRAJECTOR (denoting a central object of the scene), LANDMARK (denoting the reference entity), and SPATIAL INDICATOR (defining constraints on the spatial properties). There are three classes of spatial relations which hold between spatial markables: *REGION*, *DIRECTION* and *DISTANCE*. They are connected with the domains into which spatial relations and properties are generally grouped according to spatial information theory (Stock, 1998). SpRL-2013 introduces new spatial roles for concepts which are characteristic for motions: MOTION INDICATOR (assigned to a word or a phrase which signals a motion), PATH (denoting the path of the motion),

DISTANCE and DIRECTION (in the case of motions both used as a roles for text spans when a distance or direction is mentioned in the text). TRAJECTOR and LANDMARK are defined differently regarding a different nature of the scene (Kolomiyets et al., 2013): TRAJECTOR denotes an object which moves, starts, interrupts, resumes motion, or is forcibly involved in a motion; LANDMARK refers to a spatial context of motion.

2.3. ISO-Space 1.4

ISO-Space 1.4 (Pustejovsky et al., 2012) incorporates the annotations of static spatial information, based on SpatialML scheme, and events, borrowing from the TimeML scheme (Pustejovsky et al., 2003). Location tags (PLACE and PATH) designate regions of space, and they can both participate in spatial relationships. PATH tag is not directly related to the motion (as in the case of Spatial Role Labelling) but rather with the real or potential function of being a boundary or traversal. Non-location tags refer both to objects (SPATIAL_NE) and processes (MOTION, non-motion EVENT). There are also additional tags for relation words (SPATIAL_SIGNAL) and the tokens which capture distances and dimensions (MEASURE). The scheme is characterised by the extended set of relationship tags, which involves *QSLINK* (for topological relations), *OLINK* (for non-topological relations), *MLINK* (for distances and dimensions) and *MOVELINK* (for the representation of the path of an object in motion).

3. PST Spatial Expressions

We base our approach to modeling spatial expressions on specifications presented in Section 2. with some modifications. The most significant difference is that we use the same label (component) to annotate landmarks, trajectors, and paths, which is *spatial object*. Our motivation is that each of them refers to a physical object, and the reference to other components in the sentence defines the role of the object. Also, a single spatial object can play different roles in the same sentence, i.e., can be a trajector and a landmark at the same time. With this in mind we decided to use *landmark* and *trajector* are relations between *spatial objects* and *spatial indicators* (see Sections 3.2.3.). The other difference is the introduction of two new components: *region* (see Section 3.1.4.) and *path* (see Section 3.1.7.).
The following sections describe in detail the types of components and relations used to describe the spatial expressions. The comprehensive description of the annotation guidelines can be found in Oleksy et al. (2019).

3.1. Components

3.1.1. Spatial Object (SO)

Spatial object is a phrase denoting a material object having physical dimensions, which may be located in a three-dimensional space or in relation to which the location of another object may be described. The understanding of the *object* category in this article refers to the *object* category in the SUMO ontology (Niles and Pease, 2001), in which physical entities called *objects* are one of the subclasses of physical entities (next to *processes* and *symbols* (*content bearing physical*)). According to the authors of the mentioned ontology, the object "*corresponds roughly to the class of ordinary objects. Examples include physical objects, geographical regions, and locations of processes, the complement of objects in the physical class.*" (description of the *Object* concept in SUMO ontology). SO may function as a trajector, landmark, or path. In spatial expressions, the function of SO may be performed by nouns, pronouns, adjectives, and verbs (only if the implied subject has the status of trajector).
Examples:

- *Dom w mieście ('A house in a city')*
- *On poszedł do szkoły ('He went to school')*
- *Stoi w lesie. ('[It is] standing in the forest')*

3.1.2. Distance (DS)

Distance is a phrase denoting a distance between a trajector and landmark.
Examples:

- *Dom stoi w odległości 4 km od miasta ('The house is located at a distance 4 km from the town')*
- *Dom stoi w dużej odległości od miasta (The house is located at a great distance from the city')*

3.1.3. Direction (DR)

Direction is a phrase denoting relative or absolute direction of motion (in case of dynamic situations), or composition of localized and localizing objects (in case of static situations) described using on of three frames of reference (Levinson and Levinson, 2003).
Examples:

- *Dom jest na południe od miasta. ('The house is south of the town')*
- *Dom jest na lewo od fabryki. ('The house is on the left of the factory')*
- *Gdy miniesz most, jedź na zachód. ('When you pass the bridge, go to the west')*

3.1.4. Region (RE)

Region is a phrase denoting a part of SO. In practice, words qualified as RE have a similar function to *spatial indicator* and signal the presence of a particular relation, specifically inform that we are dealing with a part (or whole) of an object. For example, the following expressions may be qualified to the RE category: *part of*, *fragment of*, *front of*, *back of*, *side of*. A RE does not give a full, meaningful answer to the question: what is it?
Examples:

- *Fotel w tylnej części samochodu ('The seat is located in the back of the car')*
- *Dom stoi na obrzeżach miasta ('The house is on the outskirts of town')*
- *Ulica znajduje się na pograniczu śródmieścia ('The street is located on the edge of the city centre')*

3.1.5. Spatial Indicator (SI)

Spatial indicator is a phrase that signals the presence of a static spatial relationship between objects. The SI function is usually carried out by a preposition (simple or complex). This component does not have its individual meaning (synsemantic), which it acquires in combination with another component (SO).
Examples:

- *Jezioro w lesie. ('The lake in the forest')*
- *Toaleta na zewnątrz budynku. ('Toilet outside the building')*
- *Farma znajduje się na północ od muru. ('The farm is located on the north side of the wall')*
- *Drzewo rośnie w północno-zachodniej części ogrodu. ('Tree is planted in the north-western part of the garden')*

3.1.6. Motion Indicator (MI)

Motion indicator is a lexical motion exponent. Motion is a situation primarily expressed by specific motion verbs and secondarily by related nominalizations. The MI function is usually performed by verbs (or nominalizations) representing a category of verbs that denote a change in the location of an object or a change in its spatial relations with its physical environment.
Examples:

- *Pociąg jedzie z Rzymu do Wiednia. ('The train is going from Rome to Vienna')*
- *Ptak wyleciał z gniazda. ('The bird flew out of its nest')*
- *Meteoryt spadł na Ziemię. ('The meteorite fell to Earth')*

3.1.7. Path Indicator (PI)

Path indicator is a lexical exponent indicating that a spatial object (that is not a trajector) in a given situation serves as a trajector's path. The role of PI is usually assigned to prepositions of directional character: ablative (from where?), adlative (where?), perlative (which way?).
Examples:

- *Pociąg jedzie z Rzymu do Wiednia. ('The train is going from Rome to Vienna')*
- *Ptak wyleciał z gniazda. ('The bird has flew out of its nest')*
- *Meteoryt spadł na Ziemię. ('The meteorite fell to Earth')*

3.2. Relations

A spatial relation is defined as a structure based on two components – "configuration elements" (Tyler and Evans, 2003): *Trajector* and *Landmark*. They have a clearly defined (and different) denotation. It is related to the psychological concept of dividing the semantic content to be presented into figure and ground. The present approach is based on Langacker's elaboration of these concepts (Langacker, 2010). In our approach, only the SO plays a role of *trajector* or *landmark* (see Figure 1).

3.2.1. Trajector (TR)

Trajector relation occurs between a spatial object (or a region of a spatial object) and a spatial indicator (see Figure 2) or a motion indicator (see Figure 4). It indicates that the spatial object denotes a central object of the scene concerning the spatial indicator.

3.2.2. Landmark (LM)

Landmark relation occurs between a spatial object (or a region of a spatial object) and a spatial indicator (see Figure 2). It indicates that the spatial object functions as a landmark in the context of the spatial indicator. *Landmark* is a function assigned to spatial object, for which the *Trajector's* position is defined, "something a traveller stops his eye at to find his way around" (Tabakowska, 2000). *Landmark* is an object used to localize the *Trajector* (see Figure 5 for examples).

3.2.3. Argument (ARG)

For the remaining cases to describe the function of a component we use one type of a relation called *argument*. A single relation type is sufficient as its interpretation results from the component types connected with the relation (see Example#5 (Figure 5e)). All possible connections for the *argument* relation are presented on Figures 2, 3 and 4. The *argument* relation established between PI and SO has a unique status, which in some interpretations is characterized as *path* (see Example#2 (Figure 5b) or Example#4 (Figure 5d)).

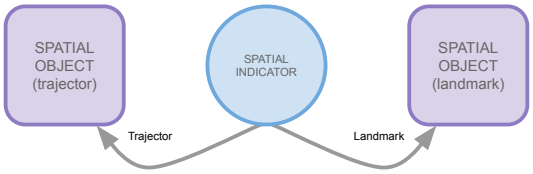


Figure 1: A structure of a spatial expression

3.3. Spatial Expressions

3.3.1. Static and dynamic situations

We distinguish two basic types of spatial expressions referring to (a) static situations and (b) dynamic situations. In both cases, phrases consist of components connected by relations. Static expressions may consist of SO, SI, DR, DS, RE (SO and SI are obligatory elements) connected by LM, TR and ARG relations. Dynamic expressions may enclose SO, MI, PI, DR, DS, RE connected by LM, TR and ARG relations. The way of establishing connections is illustrated in Figures 2, 3 and 4.

(a) **static situation** (Figure 2): two SO's and one SI; SI connected by the relation *trajector* with SO referring

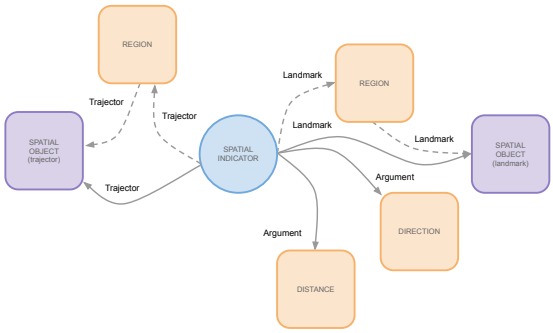


Figure 2: A structure of a static spatial expression with a spatial indicator

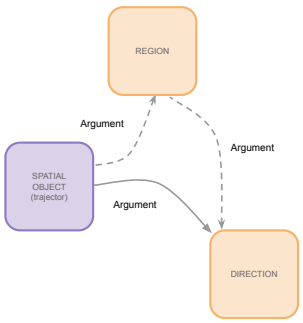


Figure 3: A structure of a static spatial expression without a spatial indicator

to the localized object and the relation *landmark* with SO denoting the physical object in relation to which the spatial relation is described. For example:

Na dole portalu są dwie romańskie rzeźby lwów z granitu (‘*There are two Romanesque sculptures of granite lions at the bottom of the portal*’)

Figure 5a presents the relations between underlined elements.

We treat as an exception the static descriptions of Direction (Figure 3). Due to the specific of Polish, SI is not obligatory if expression contains SO and the DR component. In this case, SO is linked directly to DR by the *argument* relation (or SO is linked by the *argument* relation to RE and RE has the same relationship to DR). For example:

baszta [SO] – ARG → *po prawej stronie* [DR]

‘*the tower*’ [SO] – ARG → ‘*on the right*’ [DR]

(b) **dynamic situation** (Figure 4): two SO’s components, one MI and one PI: MI is connected by the relation *trajector* with SO referring to the moving object and the relation “argument” with PI. PI is also connected by the relation “argument” with SO denoting the physical object, to which the object in motion is oriented (for example see Figure 5b).

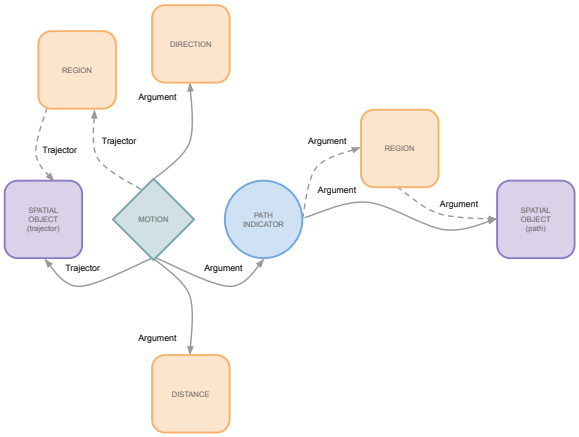


Figure 4: A structure of a dynamic spatial expression

3.3.2. Spatial Object as *Trajector* and *Landmark*

The examples given above are relatively simple and illustrate typical basic spatial expressions in PST. There are also more complex expressions in the texts, which can also be described using the annotation method. An example of a complex spatial expression is the following phrase:

> *W pobliżu ruin, kilkaset metrów dalej znajduje się drewniana kładka na potoku płynącym przez dolinę leśną [...]*
>
> *‘Near the ruins, a few hundred meters further there is a wooden footbridge over a stream flowing through a forest valley [...]’*

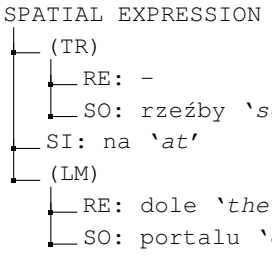
SO *potoku* (‘a stream’) plays a role of *landmark* for a SO *kładka* (‘a footbridge’) but also the role of *trajector* for another SO – *dolinę* (‘the valley’). This two expressions are illustrated on Figures 5c and 5d.

4. Corpus of Polish Spatial Texts

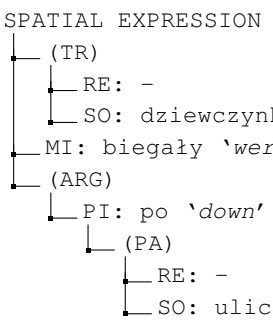
The corpus of Polish Spatial Texts (PST) was developed as a corpus for training and testing the tools for automatic recognition of spatial expressions. All documents in the corpus are manually annotated with spatial information — components and relations between them. First version of the corpus (PST 1.0) (Oleksy et al., 2018) contains the documents manually annotated with static spatial expressions. The version described in this paper (PST 2.0) was enriched with dynamic spatial expressions.

4.1. Data Source

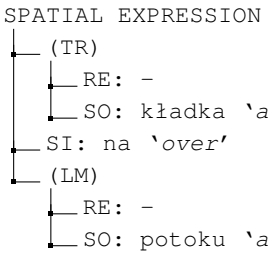
In order to provide sufficient coverage for the annotation categories occurrences in the text, the documents in the PST corpus were randomly selected from online travel blogs. Marcińczuk et al. (2016) proved that several concepts (e.g., path) rarely or never appear in the texts, which are not focused on spatial scenes description. The documents were derived from the blogs, which are published on Creative Commons license to ensure the possibility of making available free access to the final corpus.



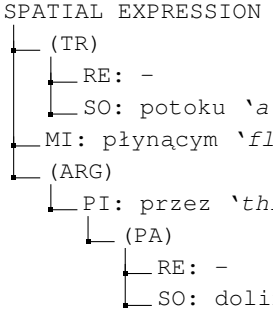
(a) Example #1: *sculptures at the bottom of the portal*



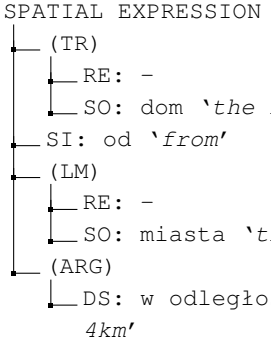
(b) Example #2: *girls were running down the streets*



(c) Example #3: *a footbridge over a stream*



(d) Example #4: *a stream flowing through a valley*



(e) Example #5: *the house at a distance 4 km from the town*

Figure 5: Examples of spatial expressions from the PST corpus

4.2. Manual Annotation

The texts collected from the sources described above were uploaded, cleaned, and pre-processed using Inforex (Marcińczuk and Oleksy, 2019), an open-source web-based system for corpora annotation[1]. The system was also the environment for manual annotation.

We annotated the corpus in several iterations — the aim was to improve annotation guidelines and data quality. During the process of annotation, inconsistencies between annotators were examined, and the guidelines were amended accordingly. We measured the inter-annotator agreement on an ongoing basis, reaching a satisfactory level.

A substantial part of discrepancies was related to the annotation extent. Linguistic phenomena that caused the most problems were: secondary prepositions, multi-word predicates, and zero (but implied) subject. Phrases with multi-word indicators were the most problematic and caused most of the disagreements.

Direction, *distance* and *region* were the concepts the most difficult to accurately identify and classify. They were at the same time the categories least represented in the corpus (see Table 2). Moreover, disagreements level was related more to the annotation incompleteness than to incorrect category assignment, e.g., most of the REGIONS annotated by annotators was accepted by the team leader, while the inter-annotator agreement was low (0.45).

Subsequently, the last iteration was performed to expand data development. A team leader resolved all inconsistencies between annotators, so the final annotations were the result of the 2+1 work procedure. Inter-annotator agreement in terms of *Positive Specific Agreement* (Hripcsak and Rothschild, 2005) calculated for all of the annotated documents in PST was 0.82 (Table 1 presents the detailed results).

Annotation category	PSA
DIRECTION	0.64
DISTANCE	0.60
REGION	0.45
MOTION INDICATOR	0.83
SPATIAL OBJECT	0.83
SPATIAL INDICATOR	0.85
PATH INDICATOR	0.83
all	0.82

Table 1: Inter-annotator agreement

The manual annotation process was divided into two parts: components labeling and tagging the relations between the components. Relations between the components were tagged when the components had been annotated, and all inconsistencies were resolved.

4.3. Corpus Statistics

The PST corpus consists of 99 annotated documents with near 12k components, 5k relations, and 2k spatial expressions. Table 2 presents detailed statistics of the PST corpus. Part A provides general statistics, including the number of

[1] https://github.com/CLARIN-PL/Inforex

documents, sentences, tokens, components, relations, and expressions. The high frequency of spatial expressions is characteristic of the corpus – there is 2 035 spatial expressions in 4 324 sentences and is related to the source of the texts (see Section 4.1.). Part B contains the number of components of specific types. The *distinct* refers to the number of unique lemmas for a given category. It is not surprising that the biggest difference between the total number and the number of unique components refers to spatial indicators – this is a relatively small number of very productive spatial prepositions. Part C presents the number of instances of relations of particular types. Parts D–F summarizes the number of relations instances between the components of particular types. For instance *region–spatial object* from part F indicates that there are 15 *argument* relations between *region* and *spatial object* (Figure 5a presents such a relation between *RE:the bottom* and *SO:of the portal*).

Tables 3, 4, 5, 6, 7 and 8 present the most frequent component values (words or phrases) for *spatial indicators*, *motion indicators*, *path indicators*, *regions*, *distances* and *directions*. The lists for *spatial indicators* and *path indicators* stand out among others, not only in terms of the number of instances. There is a fine line between the lexemes which appear most frequently and the less frequent ones. Moreover, two most frequent prepositions ("w" and "na" in the case of spatial indicators and "do" and "na" in the case of path indicators) are the key elements of the majority of the occurrences of spatial expressions (59.20% and 55.46% respectively). There is no such disproportion in the case of motion indicators, regions and distances.

4.4. License and Access

The PST corpus was released under Creative Commons license and can be obtained from CLARIN-PL Repository (Oleksy et al., 2019). The corpus is a part of the CLARIN-PL research infrastructure.

5. Summary

We have presented the process of the creation of PST corpus – Polish Spatial Texts corpus, from the conceptual stage, through the annotation process, to the results overview. As far as we know, this is the first Polish open-access corpus manually annotated with spatial information. The inter-annotator agreement related to most of annotation categories allows definite or at least tentative conclusions. We consider the corpus as a proper quality training or testing data for the systems for automatic recognition of spatial expressions. Also, statistical information on the most frequent component values provide valuable guidelines for rule-based approaches.

Future work should focus on the set of attributes for the components and relations, in order to capture such phenomena as frames of reference and type of topological relation involved. Relatively low coverage of secondary components such as *direction*, *distance* and *region* is also an issue which should be addressed in the future.

Metric	Total	Distinct
A. General		
documents	99	-
sentences	4 324	-
tokens	61 315	-
components	11 858	-
relations	5 010	-
expressions	2 035	-
B. Components		
direction	191	59
distance	74	33
motion indicator	536	186
path indicator	559	29
region	114	59
spatial indicator	1 608	47
spatial object	4 353	1 410
C. Relations		
argument	1 231	-
landmark	1 573	-
trajector	2 206	-
D. Relation *Argument* by annotation types		
motion indicator–direction	38	-
motion indicator–distance	10	-
motion indicator–path indicator	491	-
path indicator–region	23	-
path indicator–spatial object	475	-
region–spatial object	23	-
spatial indicator–direction	23	-
spatial indicator–distance	57	-
spatial object–direction	91	-
spatial object–distance	2	-
E. Relation *Landmark* by annotation types		
region–spatial object	67	-
spatial indicator–region	66	-
spatial indicator–spatial object	1439	-
F. Relation *Trajector* by annotation types		
motion indicator–region	2	-
motion indicator–spatial object	492	-
region–spatial object	15	-
spatial indicator–region	16	-
spatial indicator–spatial object	1680	-

Table 2: Statistics of the PST corpus

Acknowledgments

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

6. Bibliographical References

Chang, A., Monroe, W., Savva, M., Potts, C., and Manning, C. D. (2015). Text to 3d scene generation with rich lexical grounding. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The IAPR TC-12 benchmark – a new evaluation resource for visual information systems.

Hripcsak, G. and Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, 12(3):296–298, 05.

Jarret, A. (2019). The degree confluence project. http://www.confluence.org/ (access: 2019-11-22).

Kolomiyets, O., Kordjamshidi, P., Bethard, S., and Moens, M. (2013). SemEval-2013 Task 3: Spatial Role Labeling. Second Joint Conference on Lexical and Computational Semantics (SEM). In *Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, Atlanta, USA, june. East Stroudsburg, PA: ACL.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial Role Labeling: Towards Extraction of Spatial Relations from Natural Language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36, December.

Langacker, R. W. (2010). Reflections on the functional characterization of spatial prepositions. *Corela. Cognition, représentation, langage*, (HS-7).

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

Levinson, S. C. and Levinson, S. C. (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*, volume 5. Cambridge University Press.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S. A., and Clancy, S. (2010). SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44:263–280.

Marcińczuk, M. and Oleksy, M. (2019). Inforex — a Collaborative System for Text Corpora Annotation and Analysis Goes Open. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*, pages 711–719.

Marcińczuk, M., Oleksy, M., and Wieczorek, J. (2016). Preliminary Study on Automatic Recognition of Spatial Expressions in Polish Texts. In Petr Sojka, et al., editors, *Text, Speech, and Dialogue*, pages 154–162, Cham. Springer International Publishing.

Marge, M. and Rudnicky, A. I. (2019). Miscommunication detection and recovery in situated human–robot dialogue. *ACM Trans. Interact. Intell. Syst.*, 9(1):3:1–3:40, February.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA. ACM.

Oleksy, M., Marcińczuk, M., Bernaś, T., Wieczorek, J., and Kocoń, J. (2019). KPWr Annotation Guidelines - Spatial Expressions (2.0). CLARIN-PL digital repository http://hdl.handle.net/11321/719.

Pustejovsky, J. and Yocum, Z. (2013). Capturing motion in ISO-SpaceBank. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 25–34, Potsdam, Germany, March. Association for Computational Linguistics.

Pustejovsky, J., Castaño, J. M., Ingria, R., Saurí, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003). Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Pustejovsky, J., Moszkowicz, J., and Verhagen, M. (2012). A Linguistically Grounded Annotation Language for Spatial Information. *TAL*, 53(2):87–113.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). SemEval-2015 task 8: SpaceEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June. Association for Computational Linguistics.

Stock, O. (1998). *Spatial and temporal reasoning*. Springer Science & Business Media.

Tabakowska, E. (2000). Struktura wydarzenia w literackim tekście narracyjnym jako problem przekładu. *W: W. Kubiński, O. Kubińska, TZ Wolański (red.), Przekładając Nieprzekładalne*, pages 19–37.

Tyler, A. and Evans, V. (2003). *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge University Press.

Williams, T., Acharya, S., Schreitter, S., and Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 311–318, Piscataway, NJ, USA. IEEE Press.

7. Language Resource References

Oleksy, Marcin and Wieczorek, Jan and Bernaś, Tomasz and Marcińczuk, Michał. (2018). *Polish Spatial Texts (PST) 1.0*.

Oleksy, Marcin and Wieczorek, Jan and Bernaś, Tomasz and Marcińczuk, Michał. (2019). *Polish Spatial Texts (PST) 2.0*.

Appendices

A Most frequent components in PST 2.0

Count	SI	SI (Eng.)	%
548	w	‘in’	34.08
404	na	‘on’	25.12
82	do	‘to’	5.10
66	za	‘behind’	4.10
65	przy	‘by’	4.04
57	pod	‘under’	3.54
54	przed	‘in front of’	3.36
52	nad	‘over’	3.23
42	z	‘from’	2.61
34	od	‘from’	2.11
28	przez	‘through’	1.74
21	po	‘down’	1.31
20	u	‘at’	1.24
15	wokół	‘around’	0.93
13	ponad	‘over’	0.81

Table 3: 15 most frequent *spatial indicators* in PST 2.0

Count	MI	MI (Eng.)	%
28	schodzić	‘to get down’	5.22
20	wchodzić	‘to step up’, ‘to walk in’	3.73
19	przechodzić	‘to get across’	3.54
17	wracać	‘to return’	3.17
16	skręcać	‘to turn’	2.99
16	dotrzeć	‘to arrive’	2.99
15	iść	‘to go’	2.80
13	docierać	‘to reach’	2.43
12	jechać	‘to drive’	2.24
12	dochodzić	‘to reach’	2.24
12	wejść	‘to enter’	2.24
10	wychodzić	‘to get out’	1.87
10	pojechać	‘to drive’	1.87
9	wjechać	‘to draw in’	1.68
9	kierować	‘to drive’	1.68

Table 4: 15 most frequent *motion indicators* in PST 2.0

Count	PI	PI (Eng.)	%
215	do	‘to’	38.46
95	na	‘to’	16.99
89	z	‘from’	15.92
36	w	‘into’	6.44
29	przez	‘through’	5.19
18	po	‘through’	3.22
9	obok	‘next to’	1.61
8	od	‘from’	1.43
7	spod	‘from under’	1.25
7	nad	‘above’	1.25
6	w kierunku	‘towards’	1.07
5	w stronę	‘toward’	0.89
4	pod	‘under’	0.72
4	w górę	‘upward’	0.72
4	za	‘behind’	0.72

Table 5: 15 most frequent *path indicators* in PST 2.0

Count	RE	RE (Eng.)	%
7	koniec	‘end’	6.14
6	granica	‘border’	5.26
6	brzeg	‘edge’	5.26
5	część	‘part’	4.39
5	fragment	‘bit’	4.39
4	odcinek	‘part’	3.51
4	środek	‘middle’	3.51
3	(u) stóp	‘(at the) foot’	2.63
3	teren	‘area’	2.63
3	wnętrze	‘inside’	2.63
3	dno	‘bottom’	2.63
3	góra	‘top’	2.63
3	centrum	‘center’	2.63
3	skraj	‘brink’	2.63
3	górna część	‘top part’	2.63

Table 6: 15 most frequent *regions* in PST 2.0

Count	DI	DI (Eng.)	%
15	tuż	‘close by’	20.27
9	pobliże	‘near’	12.16
8	okolica	‘neighborhood’	10.81
4	kilka kilometrów	‘several kilometres’	5.41
3	nieco	‘slightly’	4.05
3	w oddali	‘in the distance’	4.05
2	blisko	‘near’	2.70
2	kilkaset metrów	‘a few hundred meters’	2.70
2	niedaleko	‘near’	2.70
2	pobliski	‘nearby’	2.70
2	kilkadziesiąt centymetrów	‘a few tens of centimetres’	2.70
1	200 metrów	‘200 meters’	1.35
1	zaraz	‘around’	1.35
1	na wyciągnięcie ręki	‘at your fingertips’	1.35
1	przez kilka metrów	‘for a few meters’	1.35

Table 7: 15 most frequent *distances* in PST 2.0

Count	DR	DR (Eng.)	%
18	w dół	‘down’	9.38
14	z lewej	‘on the left’	7.29
13	z prawej	‘on the right’	6.77
11	w lewo	‘left’	5.73
10	w górę	‘up’	5.21
9	w prawo	‘right’	4.69
7	po lewej	‘on the left’	3.65
7	na zachód	‘to the west’	3.65
7	na prawo	‘to the right’	3.65
6	po prawej	‘on the right’	3.13
6	na południe	‘to the north’	3.13
5	na wschód	‘to the east’	2.60
5	z tyłu	‘back’	2.60
4	po drugiej stronie	‘on the other side’	2.08
4	na wprost	‘in front of’	2.08

Table 8: 15 most frequent *directions* in PST 2.0