SpaceRef: A Corpus of Street-Level Geographic Descriptions

Jana Götze and Johan Boye
School of Computer Science and Communication
KTH Royal Institute of Technology
Stockholm, Sweden
{jagoetze,jboye}@kth.se

Abstract

This article describes SpaceRef, a corpus of street-level geographic descriptions. Pedestrians are walking a route in a (real) urban environment, describing their actions. Their position is automatically logged, their speech is manually transcribed, and their references to objects are manually annotated with respect to a crowdsourced geographic database. We describe how the data was collected and annotated, and how it has been used in the context of creating resources for an automatic pedestrian navigation system.

Keywords: reference resolution, corpus, pedestrian wayfinding

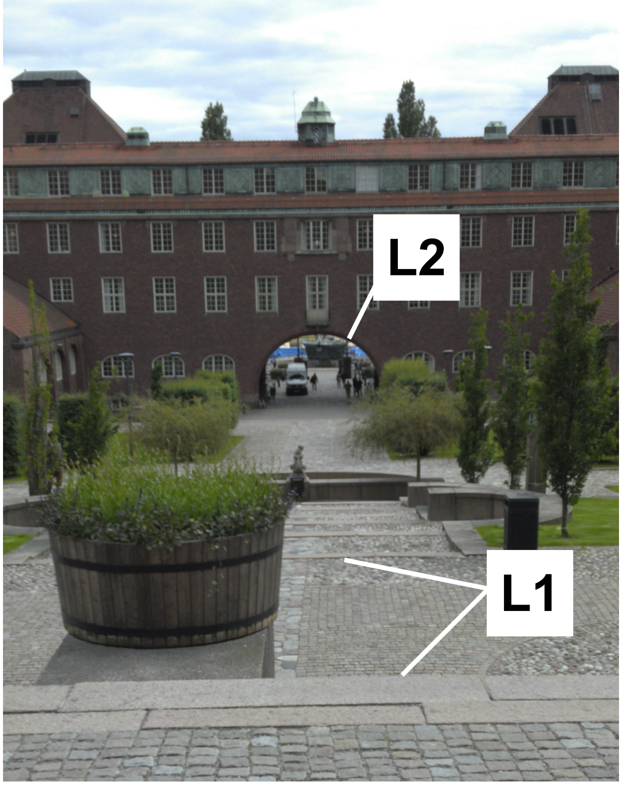
1. Introduction

We are introducing SpaceRef, a small corpus of spoken geographic and spatial descriptions given by pedestrians while moving in an urban environment. The corpus contains transcribed utterances, along with the pedestrians' GPS coordinates and information about which objects in their geographical surroundings they are referring to. We believe this corpus will be a useful resource for researchers studying reference resolution, landmark salience in the context of route instructions (Richter, 2013), geographical dialogue systems (Boye et al., 2014), and qualitative spatial reasoning (Freksa, 1991).

These research problems are tightly interconnected, but often studied in isolation and on the basis of different data sources. References to objects in the surrounding physical environment, as they are often found in situated speech, are typically studied in small-scale environments such as objects aligned on a table (Matuszek et al., 2014). Mechanisms that model human-like route instructions, including references to landmarks, are often based on studies where participants give written instructions or for prospective routes, i.e. instructions that need to be remembered by the route follower (). One reason for this is that collecting data in real environments is generally a time-consuming undertaking that offers limited possibility to control the experiment conditions.

In this paper, we describe our efforts of collecting data of pedestrians walking in an urban environment and describing the actions that they are doing as if talking to a route follower. The data has been collected with two specific applications in mind, namely as basis for developing an algorithms that resolves references in the real physical environment while a pedestrian is moving in it, and as basis for deriving models of landmark salience from mentions of landmarks in specific routing situations. Both applications are necessary in a system that can automatically give route instructions to pedestrians (Boye et al., 2014). The aim was therefore to put the pedestrians into a situation that resembles as closely as possible the situation of a potential user of such a system.

In the remainder of this paper, we describe the two studies that were the reason for collecting the data (Section 2) and



'I continue in a southwesterly direction *down the steps* [L1] *towards the arch at the bottom* [L2]'
Excerpt from its representation in SpaceRef (the identifier numbers for L1 and L2 are retrieved from OpenStreetMap):

```
utterance : "i continue in a southwesterly
             direction down the steps
             towards the arch at the bottom"
time      : '2:1:10:14:41:8571'
latitude  : "59.34787"
longitude : "18.07406"
RE : "the steps"
     id="1" referent id="20680216"
RE : "towards the arch at the bottom"
     id="2" referent id="163195369"
```

Figure 1: Example utterance and schematic SpaceRef representation

how the data was collected and annotated for the purpose of these two studies (Section 3). Section 4 discusses related work in terms of similar corpora and annotation schemes. Section 5 discusses open questions such as further potential uses for this data.

2. Corpus Usage

The SPACEREF data was collected in the context of developing a system that can give automatic and interactive spoken route instructions to pedestrians (Boye et al., 2014). Two required functionalities of such a system are choosing appropriate landmarks to incorporate into route instructions and resolving the pedestrian's references to objects in the environment.

Spoken language often contains references to objects in the surrounding physical environment. Any theory or computer system that endeavors to interpret spoken language therefore needs a mechanism for *resolving* such referential expressions, i.e. linking linguistic expressions to entities in the external world.[1] In any moderately complex environment, there will often be plenty of entities that can be targets of a particular referential expression. In order to find the right entity it is therefore important to assess and take into account how *salient* each object is in a particular situation. A reliable salience estimate for geographical situations can in its turn be used by a way-finding geographical system for selecting appropriate landmarks on which to base route instructions. A geographical spoken dialogue system must be able to both interpret and generate utterances containing references to real-world objects in the environment.

2.1. Landmark Salience

Landmarks play a vital role in pedestrian wayfinding, both when giving and when understanding route instructions (Denis, 1997; Lovelace et al., 1999). Picking an appropriate landmark for route instructions is a difficult task and is usually based on heuristics about what makes objects salient (Raubal and Winter, 2002). Pedestrians in SPACEREF are following a given route, describing their actions as they are walking. They are perceiving the environment directly, in the same way as potential users of an automatic system and extensively refer to landmarks. The aim was to learn from the pedestrians' uses of landmarks.

We have used this corpus to compute models that can predict landmark salience (Götze and Boye, 2013; Götze and Boye, 2015a). The models are based on the observation that every time the pedestrian is choosing a landmark to describe his path, he is preferring that landmark over all other objects in the vicinity. We trained a Support Vector Machine model that ranks all objects in the near vicinity according to these user preferences, and found that this model generalizes well to new unseen situations, i.e. the model is able to predict to a large extent which landmark the user would prefer to use in a description in a new situation.

For this task, SPACEREF gives information about what landmarks were referred to at what location, what other objects could have been referred to, and where the pedestrian was headed (but not how the pedestrian phrased their reference).

2.2. Reference Resolution (RR)

Recently, there has been increased interest in resolving references to real-world entities, e.g. in the context of Human-Robot Interaction (Matuszek et al., 2014) and grounding language using non-linguistic information (Iida et al., 2011; Kennington and Schlangen, 2015). However, most of this research studies the problem in laboratory settings, where subjects refer to a small set of known objects.

We are currently studying how the SPACEREF data can be used for reference resolution. An initial study on part of the data was presented in (Götze and Boye, 2015b), and a more extensive study has been completed (Götze and Boye, 2016). Resolving references in this large-scale environment is a complex task, differing from the small-scale tabletop or screen settings referenced above, where all available objects are visible at once and all their relevant properties such as size and color are known. By contrast, we are continuously and automatically updating the list of nearby objects together with their properties, the *candidate set*, from the city model on the basis of the current user position.

For this task, SPACEREF gives information about what object a pedestrian referred to at what location, what other objects could have been referred to, and how the pedestrian referred to the object (or objects).

Using this corpus for reference resolution introduces a number of additional sources of noise from both the user data and the GIS database that need to be addressed:

1. Pedestrians are moving while they are describing, meaning that the set of objects they can see changes continuously and needs to be recomputed for each new utterance. This computation is currently done on the basis of the pedestrian's position. This latitude/longitude position from the GPS data is however imperfect and it can therefore happen that the referenced object is not part of the candidate set even though it is in the city model.

2. As we have described in (Götze and Boye, 2015b), it is not always obvious how many objects constitute the correct referent of a RE. For example, street intersections of larger streets typically consist of more than one node.

3. We are currently working on manually transcribed speech. Introducing RR to an automatic system means resolving REs on the basis of speech recognition results, which is likely to contain a higher number of errors because of background noise from the street.

3. Corpus Description

The SPACEREF corpus contains transcribed user speech that is annotated with the pedestrians' position and information about which objects they are referring to, as exemplified in Figure 1. For geographical information, we

[1] Note, incidentally, how such *reference resolution* differs from anaphora resolution where the aim is to find co-referring expressions in text (but disregards the problem of finding the entity being referred to).

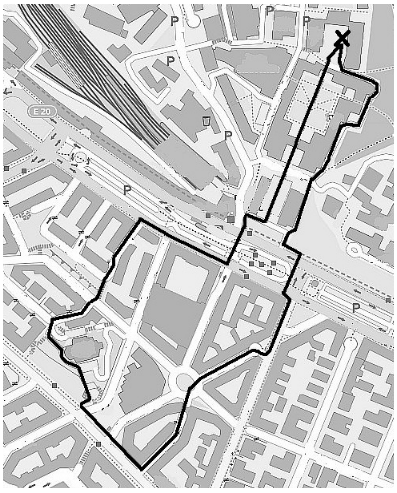


Figure 2: The map of the route that participants were walking © OpenStreetMap contributors

are relying on the Openstreetmap (OSM) database[2] (Haklay and Weber, 2008). When humans express knowledge about wayfinding they frequently use landmarks (Denis, 1997; Lovelace et al., 1999; Denis et al., 1999), which is reflected in SpaceRef.

3.1. Participants, Task, and Setup

This data was collected in a Wizard-of-Oz dialog setting (Dahlbäck and Jönsson, 1989). The 10 participants were instructed to walk a given path and describe their actions to the system in a way that would make it possible for the system to follow them without knowledge of their position. The participants were given an unlabelled map that contained no names or common symbols in order to force them to rely on perceptual information. The map is shown in Figure 2, where start and end point are marked with 'X'. Participants decided themselves in which direction to walk the round tour.

The role of the wizard was to acknowledge the participants' instructions and interfere only when an instruction was either unintelligible because of background noise from cars or when it was obviously ambiguous or wrong (e.g. asking for clarification when the participant confused *left* and *right*).

The participants' speech as well as positions as GPS coordinates were collected by means of an Android phone (Motorola Razr) application. The wizard was sitting in the lab and using an interface where he could see the walking participant's position and send text to the phone that was read out to the participant through the phone's text-to-speech application (Hill et al., 2012).

3.2. Data and Annotation

A summary of the participants as well as the corpus size in terms of REs can be found in Table 1. Most participants spend their working days on the university campus from where the route started and are therefore familiar with the immediate surroundings of the campus. Another part of the route led through a residential area that the participants visited rarely or never. For the overall route, they reported to be familiar with the area (4.4 on a scale from 1 –"not familiar at all" to 6 –"very familiar").

Each participant's data comes as an xml file containing the segmented and transcribed speech. For each speech segment, the participant's GPS coordinates as well as REs that refer to objects in the environment are annotated with either the OSM ID(s) from the city model, or an indicator that the corresponding object is not mapped in the database ('nm') or the referent is unknown ('unk').

The two tasks, deriving salience models and resolving references, require different levels of annotation regarding what constitutes a referring expressions. For the task of deriving salience models we only require to know what object has been mentioned at what location, it is not necessary to know exactly what words the pedestrian used. The positional information (GPS coordinates) that is associated with an utterance is the pedestrian's position at the beginning of the utterance. We use this position to determine the candidate set of objects and all objects that are mentioned in this particular utterance are annotated as a landmark, i.e. with the OSM ID of the object(s).

For the task of reference resolution we require to know exactly what words the pedestrian used to refer to an object. Typically, references to physical objects are expressed with noun phrases. As described in Section 2.2, we want to adopt the words-as-classifiers approach described in (Kennington and Schlangen, 2015). This approach trains a classifier for each word in a referring expression, based on features that describe the candidate objects: information about their position, their type, and their relation to each other. When applying the classifiers to a new referring expression, each word determines whether the expression can refer to an object in the new candidate set. Intuitively, we expect that more information than just the noun phrase will contribute to the correct resolution of a referring expression. For example, the classifier for the preposition *along* will learn to associate itself with objects of type `street` or `building`, but not with type `shop`. Therefore, we define a RE rather loosely as any substring from the utterance that contains information about an object. Specifically we included spatial prepositions like *along* and *through*, transitive motion verbs like *cross*, and mentions of relative direction like *to the left*.

These are some examples:

- *"There's a fountain in the middle of the park"*
- *"I'm now walking through the trees towards the road"*
- *"Right so on my left there's a green fence which is pointy at the top"*
- *"Okay so now I'm going down towards the bigger road"*

Additional data includes the full GPS path for each participant, the OSM city model file, a file containing a specification of which street segments constitute one street, and participants' answers from questionnaires that they answered after they carried out the task. The speech was originally

[2] www.openstreetmap.org

Table 1: Summary of the SpaceRef data

Number of participants	11
Male/female participants	9 / 2
Average age	27.4
Familiarity with the area (1^- to 6^+)	4.4
Voice application usage (1^- to 6^+)	1.8
Walking time	5h44m
Utterances	1,676
Referring expressions	1,323
Unique referents per participant	54.7
REs without referent on the map	58 (4%)

Number of nodes	29,451
Number of ways	4,031
Number of unique tags	5,142
Number of unique tag keys	210
Average number of objects in candidate set	33

Table 2: Summary of the Openstreetmap data

transcribed using the Higgins Annotation Tool[3] and is available in this format without annotation.

What is not annotated are other REs such as personal pronouns that do not refer to objects in the environment, "negative" references, such as "there is no intersection", as well as events or actions.

3.3. Openstreetmap (OSM)

Openstreetmap (OSM) is a crowd-sourcing project that creates maps of the world. The data is available under the Open Data Commons Open Database Licence (ODbL) and has been extensively used for research of various kinds such as navigation (Hentschel and Wagner, 2010) and education (Bartoschek and Keßler, 2013). Especially in urban areas, the map coverage is high, making it suitable for our purpose of urban pedestrian navigation, where users refer to a variety of objects. Table 2 shows that for our study area, only about 4% of all REs refer to objects that are not represented on the map.

Openstreetmap represents objects as two different data types: *nodes* and *ways*. Each node is described by its latitude/longitude coordinates and ways are described by the nodes that make up a street, a building, or an area. Each object has an ID and can be tagged with key-value pairs expressing a wide range of information about the object. Contributors are asked to adhere to an extensive wiki specification (`wiki.openstreetmap.org/wiki/Map_Features`). For example, in Figure 1, object L2 ("the arch") is represented as follows:

```
<way id="163195369">
<nd ref="1749442658"/>
<nd ref="1749442656"/>
<tag k="highway" v="footway"/>
<tag k="layer" v="-1"/>
<tag k="source" v="yahoo; survey"/>
<tag k="tunnel" v="yes"/>
</way>
```

4. Related Work

The PURSUIT corpus (Blaylock, 2011) is most similar to SpaceRef in that it contains both GPS tracks and annotated mentions of spatial entities. PURSUIT is different in that it was collected from car drivers and is annotated with respect to two different GIS databases, which are reported to cover 82.5% of the geographical mentions. The entities are classified into one of four classes (streets, intersections, addresses, other locs), and identified by name and/or lat/lon coordinate. However, the PURSUIT annotations do only contain information about which objects were referred to but neither are the properties of these objects known nor what other objects are available for reference at each position, thus making it insufficient for the task of reference resolution.

Another similar dataset that is used for the task of reference resolution is presented by Misu et al. (2014). Like the PURSUIT corpus, this is a collection of car drivers moving through an urban environment. Instead of describing their environment, the participants in this data collection pose queries about Point of Interest (POI) to an in-car dialog system. The data contains speech and GPS information like SpaceRef, and additionally information about the driver's head pose. However, the information about the POIs is manually annotated.

As mentioned in Section 1, several studies have collected data in small-scale environments, such as objects on a tabletop (Matuszek et al., 2014; Kennington and Schlangen, 2015) or on a computer screen (Iida et al., 2011; Funakoshi et al., 2012). Some studies have also worked with virtual environments in which all object properties are known (Schütte et al., 2010).

SpatialML (Mani et al., 2008) is an annotation scheme for geographical place mentions in natural language. In contrast to our annotation, each subpart of a mention is tagged with an own tag, and different kinds of relations are explicitly distinguished. Geographic entities are annotated with respect to a certain gazetteer, similar to our OSM annotation. The annotation currently permits only one entity with a corresponding lat/lon specification to be annotated. However, for the purpose of street-level navigation it is useful to know when an object has a larger extension, as is the case for buildings or areas.

An annotation scheme that would be better suited is ISO-Space (Pustejovsky et al., 2011), as it extends SpatialML to account for a wider range of spatial expressions. Although the framework was set up with written text in mind, it should be possible to apply to transcriptions of spoken route directions. We leave the investigation of how such annotation schemes can be integrated with this data for future work.

Finally, newspaper texts and travel reports also contain spatial references, but typically on a more coarse-grained level,

[3] `http://www.speech.kth.se/hat/`

e.g. to cities or countries. In the recent SpaceEval task (Pustejovsky et al., 2015), the goal is to automatically find and classify the relevant parts of spatial referring expressions (rather than resolve them).

5. Discussion and Open Questions

The SPACEREF corpus introduces data that can, among other things, be used for situated reference resolution. There are a number of questions and extensions to be addressed in future work:

All pedestrians are walking the same path (and within two weeks' time), making their references comparable, both in what they refer to and how they refer to it. We would like to encourage similar data collections in different (urban) areas, in English as well as other languages, with possibly varying tasks. In SPACEREF, pedestrians are not carrying out any particular task and the tour has the nature of a walk without any particular goal.

Annotation of this data was laborious, using Openstreetmap's online interface. The crowd-sourced nature of the OSM data should make it possible to integrate geographic annotation into existing annotation tools, such as the NITE XML toolkit (Carletta et al., 2005).

Currently, only mentions of geographic objects are annotated. The corpus can potentially also be used for analysis of action descriptions, similar to the PURSUIT corpus (Blaylock and Allen, 2008; Blaylock et al., 2009). Applying the ISO-Space framework as mentioned in Section 4 and using the corpus to test the framework's expressiveness for such situated language use appears to be a useful addition for studies of spatial language.

Finally, data like this is potentially useful to extend the geographic database with additional information, e.g. tags about details such as color, material, or accessibility, similar to (Meena et al., 2014).

6. Conclusion

We have described SPACEREF, a corpus of pedestrians describing their way while walking. We believe that this corpus is a useful addition to studying the problem of grounding language in the real world and would like to encourage more such "out-of-the-lab" data collections.

7. Bibliographical References

Bartoschek, T. and Keßler, C. (2013). VGI in Education: From K-12 to Graduate Studies. In *Crowdsourcing Geographic Knowledge*, pages 341–360.

Blaylock, N. and Allen, J. (2008). Real-time path descriptions grounded with gps tracks: a preliminary report. In *LREC Workshop on Methodologies and Resources for Processing Spatial Language*, pages 25–27.

Blaylock, N., Swain, B., and Allen, J. (2009). Mining geospatial path data from natural language descriptions. In *Proc. of the 1st ACM SIGSPATIAL GIS International Workshop on Querying and Mining Uncertain Spatio-Temporal Data*.

Blaylock, N. (2011). Semantic annotation of street-level geospatial entities. In *Proc. of the IEEE ICSC Workshop on Semantic Annotation for Computational Linguistic Resources*.

Boye, J., Fredriksson, M., Götze, J., Gustafson, J., and Königsmann, J. (2014). Walk this way: Spatial grounding for city exploration. *Natural Interaction with Robots, Knowbots and Smartphones*, pages 59–67.

Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The NITE XML Toolkit: Data Model and Query Language. *Language Resources and Evaluation*, 39(4):313–334.

Dahlbäck, N. and Jönsson, A. (1989). Empirical studies of discourse representations for natural language interfaces. In *Proc. of EACL*, pages 291–298.

Denis, M., Pazzaglia, F., Cornoldi, C., and Bertolo, L. (1999). Spatial discourse and navigation: an analysis of route directions in the city of Venice. *Appl. Cogn. Psych.*, 13(2):145–174.

Denis, M. (1997). The description of routes: A cognitive approach to the production of spatial discourse. *Current Psychology of Cognition*, 16(4):409–458.

Freksa, C. (1991). Qualitative spatial reasoning. In *Cognitive and Linguistic Aspects of Geographic Space*, volume 63, pages 361–372.

Funakoshi, K., Nakano, M., Tokunaga, T., and Iida, R. (2012). A unified probabilistic approach to referring expressions. In *Proc. of SIGdial*, pages 237–246.

Götze, J. and Boye, J. (2013). Deriving salience models from human route directions. In *Workshop on Computational Models of Spatial Language Interpretation and Generation 2013 : (CoSLI-3)*, pages 36–41.

Götze, J. and Boye, J. (2015a). Learning Landmark Salience Models from Users' Route Instructions. Presented at the Symposium on Location Based Services, extended version submitted to Journal of LBS.

Götze, J. and Boye, J. (2015b). Resolving spatial references using crowdsourced geographical data. In *Proc. of NODALIDA*, pages 61–68. Linköping University Electronic Press.

Götze, J. and Boye, J. (2016). Reference resolution in spatially-aware systems. submitted to ACL.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *Pervasive Computing*, 7(4):12–18.

Hentschel, M. and Wagner, B. (2010). Autonomous robot navigation based on OpenStreetMap geodata. In *Proc. of ITSC*, pages 1645–1650.

Hill, R., Götze, J., and Webber, B. (2012). Final data release, Wizard-of-Oz (WoZ) experiments. `http://www.spacebook-project.eu/pubs/D6.1.2.pdf`.

Iida, R., Yasuhara, M., and Tokunaga, T. (2011). Multimodal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proc. of IJCNLP*, pages 84–92.

Kennington, C. and Schlangen, D. (2015). Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proc. of ACL-IJCNLP*, pages 292–301.

Lovelace, K. L., Hegarty, M., and Montello, D. R. (1999). Elements of good route directions in familiar and unfamiliar environments. In *Spatial Information Theory.*

Cognitive and Computational Foundations of GIS, volume 1661, pages 65–82.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). SpatialML: Annotation Scheme, Corpora, and Tools. In *Proc. of LREC*.

Matuszek, C., Bo, L., Zettlemoyer, L., and Fox, D. (2014). Learning from unscripted deictic gesture and language for human-robot interactions. In *Proc. of AAAI*, pages 2556–2563.

Meena, R., Boye, J., Skantze, G., and Gustafson, J. (2014). Crowdsourcing street-level geographic information using a spoken dialogue system. In *Proc. of SIGdial*, pages 2–11.

Misu, T., Raux, A., Gupta, R., and Lane, I. (2014). Situated language understanding at 25 miles per hour. In *Proc. of SIGdial*, pages 22–31.

Pustejovsky, J., Moszkowicz, J. L., and Verhagen, M. (2011). Iso-space: The annotation of spatial information in language. In *Proc. of the 6th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). SemEval-2015 Task 8: SpaceEval.

Raubal, M. and Winter, S. (2002). Enriching wayfinding instructions with local landmarks. In *Geographic Information Science*, pages 243–259.

Richter, K.-F., (2013). *Cognitive and Linguistic Aspects of Geographic Space: New Perspectives on Geographic Information Research*, chapter Prospects and Challenges of Landmarks in Navigation Services, pages 83–97.

Schütte, N., Kelleher, J., and Mac Namee, B. (2010). Visual salience and reference resolution in situated dialogues: A corpus-based evaluation. In *AAAI Symposium on Dialog with Robots*.