Resolving Spatial References using Crowdsourced Geographical Data

Jana Götze and Johan Boye
KTH, School of Computer Science and Communication
100 44 Stockholm, Sweden
jagoetze,jboye@csc.kth.se

Abstract

We present a study in which we seek to interpret spatial references that are part of in-situ route descriptions. Our aim is to resolve these references to actual entities and places in the city using a crowdsourced geographic database (OpenStreetMap). We discuss the problems related to this task, and present a possible automatic reference resolution method that can find the correct referent in 68% of the cases using features that are easily computable from the map.

1 Introduction

When humans give route instructions to each other, such instructions typically involve a wide range of references, such as references to landmarks ("Turn at the church."), to the spatial configuration ("The road is bending to the left."), to the current path of movement ("Keep walking along this road."), or to the direction of movement ("You should turn to the right."). Determining which places and objects are referred to is a significant part of designing geographical information systems that aim at interacting with the user in natural language. A long-term goal for our automatic navigation system (Boye et al., 2014) is to be able to ground that a route instruction was understood or to enable the user to ask questions about a particular landmark. This requires resolving the user's geographic references.

Resolving referring expressions (REs) to entities in the world is an ongoing area of research.[1] In written text, including web pages and search queries, references are often to geographic entities such as cities or countries (Amitay et al., 2004; Martins et al., 2006; Pouliquen et al., 2006). In spoken language, the domain is typically restricted to a task that one or more speakers are solving by referring to the objects that are involved, e.g. the pieces of a puzzle (Funakoshi et al., 2012; Matuszek et al., 2014).

This paper addresses the problem of mapping from linguistic REs that refer to aspects of space to objects in a map representation of that space. We collected a number of path descriptions from pedestrians, similar to the corpus of (Blaylock, 2011). The REs we are interested in refer to entities in a real urban environment and the map representation is general rather than tailored to this particular problem. We give an overview of the kinds of knowledge needed to resolve different kinds of references that speakers use to describe their environment while navigating in it. We discuss the challenges that occur when real language data meets real spatial data and suggest ways to address them.

2 Representing Space: OpenStreetMap

OpenStreetMap (OSM) is a crowdsourcing project that creates a geographical knowledge base (Haklay and Weber, 2008). Similar to Wikipedia, the data is open[2] and has been used for research projects in different areas, as well as for education and to create maps for special needs, such as bicycle or hiking maps.[3]

The geographic data can be downloaded in an xml format, Figure 1 shows a short extract. There are two basic data types that are used to represent objects in the OSM database: *nodes* and *ways*. *Ways* are sequences of *nodes*, used for representing a wide variety of objects, such as roads,

[1] Note that this is different from coreference resolution, where the objective is to identify those expressions in a text that refer to the same entity, but not to identify what that entity is (Mitkov, 2010).

[2] http://www.openstreetmap.org/copyright

[3] For an overview of OSM-based applications for research, education, and other purposes, cf. http://wiki.openstreetmap.org

```
<node id="485981500" lat="59.3360310" lon="18.0510617">
  <tag k="amenity" v="bench"/>
</node>
<node id="674212016" lat="59.3380430" lon="18.0529256">
  <tag k="addr:housenumber" v="15"/> <tag k="addr:street" v="Upplandsgatan"/>
</node>
<way id="39228957">
  <nd ref="469951578"/> <nd ref="469955649"/> <nd ref="469952066"/>
  <tag k="highway" v="footway"/> <tag k="surface" v="paved"/>
</way>
```

Figure 1: An extract of OpenStreetMap data. Each entity has an ID and can be annotated with several tags. This extract shows two *nodes* (a bench and a street address), and a *way*, consisting of several nodes.

squares, areas and buildings (in the three latter cases, the first node in the sequence is the same as the last node, and hence the way forms the perimeter of a polygon). An intersection between two streets is represented by the node where the ways corresponding to the streets meet. Both nodes and ways can be annotated with a set of tags to specify names and types, and additional information such as opening times or links to homepages.

The OSM wiki explains the available set of tags[4] and how they should be used. However, the geographical situation is often not as clear as the given examples and the same kind of object can be represented in different ways, as we will describe further in Section 5. Furthermore, the data is also incomplete: Not all things that speakers mention are mapped, not all details about entities are mapped, and there are errors, e.g. spelling mistakes or wrong tags.

On the other hand, OSM often provides a fine level of detail in urban areas for objects that can be useful for pedestrian navigation. This includes information about many kinds of landmarks and smaller objects such as artworks or benches. The crowdsourced nature of the data also makes it possible for the crowd to correct mistakes in spellings or positions, as well as to keep the map updated.

3 Spatial Descriptions

In order to obtain REs that are used while the speaker is moving in the environment on foot, we carried out the following study.

3.1 Data Collection

For this study, we used data from a previous data collection (Götze and Boye, 2013) in which subjects were asked to walk a specific route and describe their path in a way that would make it possible for someone to follow them. We thereby put participants into the same environment in which we would later like to guide them. Instead of reading from a 2-dimensional map, our participants can now see the environment in the same way as users of a route-giving system experience it.

The experiment was set up as a Wizard-of-Oz situation in which the participants were asked to describe to a spoken dialog system with the task of making it understand. They were told that the system, like them, had a 3-dimensional and 1st-person view of the environment. The participants were not instructed to interact with the system in any special language but were advised to try out what they thought was suitable and that the system would ask them if it needed clarification, in which case they should stop until the situation was clarified. In this way, the experimenter was able to interfere in situations where an instruction was evidently ambiguous. Otherwise, the experimenter took as little initiative as possible in order to avoid influencing them in their choice of REs.

The data was collected in English,[5] in which all participants reported to be fluent. All were slightly familiar or familiar with the area and all were able to complete the task.

The route that the participants were asked to walk was a round tour that started and ended outside the doors of our laboratory. The route was approximately two kilometers long and was given

[4] http://wiki.openstreetmap.org/wiki/Map_Features

[5] The data collection was carried out as part of the European Spacebook project: www.spacebook-project.eu

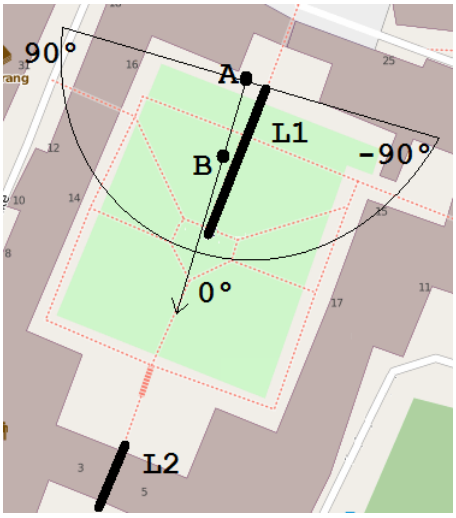

Figure 2: An example segment for the utterance: "I continue in a this direction down *the steps* [L1] towards *the arch* [L2]" A and B indicate the start and the goal position respectively. The lines indicate the speaker's direction and field of view.

to the participants on an unlabelled map. The map had street and other names removed, as well as common symbols, e.g. for churches or bus stops.

The recorded speech was transcribed and segmented into utterances, and aligned with the GPS signal. Figure 2 shows an example utterance, the GPS coordinates (the points A and B) indicate where the instruction was given and where the next instruction followed. In this example, the participant referred to two objects, "the steps" and "the arch". Both of these objects are OSM ways and indicated by the lines L1 and L2 in the figure.

Here, we consider the route descriptions of three of the study participants. Note that none of the descriptions contain any names of streets because we asked participants to avoid them. The original purpose of collecting this data was to investigate what landmarks are used for guiding someone and street names are known to be hard to recognize in a route finding scenario (Tom and Denis, 2004). We are extracting all REs they used, but restrict ourselves here to noun phrases that refer to entities that could in principle be represented on a map (explicitly or implicitly), such as "a junction" or "the church". Noun phrases that refer to directions ("to *the left*") or that are referring to the task ("I made *a mistake*") are excluded. This results in a total of 398 REs, 150 by participant *A*, 122 by participant *B*, and 126 by participant *C*.

3.2 Common Referring Expressions

Many REs (ca. 97%) contain the **type** of the entity as interpreted by the describer, e.g. "a small tunnel", "the parking lot", "the street ahead".

Names, e.g. "Baldersgatan", "Engelbrektsskolan", "the Algerian embassy", can occur in REs, usually for streets or for objects whose names are clearly visible. In our data, the describers use names in 2–15% of the REs.

In around 3–9% of the REs in our data the description is more detailed and specifies a certain **part of an entity**, e.g. "the middle of the park", "an entrance to the station", "the end of the road".

A RE includes the object's **location relative** to the speaker in around 27% of the cases, e.g. "a fountain to my left", "ahead of me is the bus station", "on the right hand side of the building", "a building to my right".

Plurals and sets, e.g. "some steps", "a collection of trees", can occur in the REs. Several objects can be referred to as one or one object can be perceived as many.

Some references (ca. 3%) describe **topographical features** of the terrain, e.g. "the hill", "a slight incline", "the arch at the bottom".

4 What we Need to Resolve Spatial References

We can now look at the different kinds of information that we need to resolve the example references and check whether this information is in principle inferrable from the OSM geographical representation.

4.1 Types of Knowledge Needed

Position, distance, and angles

We need to know the placement of objects on the map as well as the speaker's current and previous position to determine distances and relative directions. For example, in expressions like "I'm walking toward the street." where we want to exclude entities that are behind the speaker.

Visibility

In our dataset, speakers are describing the way they are walking and we can therefore assume that they are referring to objects they can see. This assumes knowledge about the height and extension of objects as well as topographical knowledge to know whether the speaker or an object is located on e.g. a hill.

Type information

Most often, objects are referred to by their type. Describers can use different expressions to refer to the same type: "I am crossing the *street/road*", and describers can use the same expression to refer to different types: A *street* could also be a bike lane or a footway. Information about how types are related to one another as well as which expressions can designate which types in the map is needed to resolve such ambiguities.

Names

Although not many of the REs in our corpus contain names, they can be useful to reduce the number of possible referents. A method is needed to map colloquial or shortened names to those in the database, as well as to resolve ambiguities where several entities have the same name, e.g. a bus stop may be named after the hospital where it is located.

Topography

In order to resolve REs that refer to topographical features, knowledge about elevation is needed.

Discourse history

We are dealing with continuous descriptions and speakers who are moving through the environment as they are speaking. Speakers are referring to some objects several times, e.g. to describe them in more detail. This results in the use of pronouns and short descriptions that we can only resolve by taking into account previous utterances (as well as already found referents):

Position	Utterance
P_t	"So I'm right in front of *the arcs*."
P_{t+1}	"and I'm walking through *them*."

4.2 When to Reject a Solution

No map of a real urban environment can be assumed to be complete. We therefore need a mechanism to decide that we cannot resolve the reference to anything in the map representation. This can be decided on the basis of e.g. distance, visibility, and type. If the describer is talking about a pedestrian crossing, and there is none within a small radius, we can reject the expression as unresolvable. If the describer is talking about a building, it might be visible from further away and we can extend the radius to look for possible referents.

4.3 Using OpenStreetMap

Let us now consider how we can obtain this kind of knowledge from OpenStreetMap (OSM). Recall that we are assuming knowledge about the speaker's position.

Knowledge that can be obtained directly

Recall from Section 2, that OSM entities (*nodes* and *ways*) are tagged with their **position** in terms of latitude and longitude, as well as information about their **type** and their **name** (cf. Figure 1).

Information about **topography** is in principle possible to obtain from OSM. The tag `incline` can be used to specify the steepness of a way. The tags `natural` and `ele` can be used to specify a peak and a point's elevation above sealevel. To specify the height of buildings, OSM provides the tag `height`. However, these topographical tags are rarely used in the urban environment that corresponds to the REs from our data.

Knowledge that can be inferred

Both **distance** and **angles** can easily be inferred using the speaker's and the entities' positions. As mentioned above, the concept of an *intersection* can be inferred by checking how many streets (or OSM ways) are meeting in a node. If more than two streets meet, we can assume that the node is a junction. This knowledge is needed for descriptions that specify a certain part of a street, such as "the end of the street" or "the corner of street X and street Y".

Information about **visibility** can be computed from knowledge about topography and distance if it is available. In order to approximate knowledge on visibility where it is not available, we can check whether there is a free line of sight from the speaker to an entity, i.e. whether there is a building in between the speaker and the entity.

Some **types** do not have to be explicitly represented in the form of tags, but can be inferred. For example, in order to determine which buildings make up a university campus or a hospital complex, it may be possible to group them on the basis of their name.

4.4 Other Sources of Knowledge

When speakers describe something by its type ("I can see *a fountain*."), then this type does not necessarily correspond to the type as used in OSM. For example, what describers call a "street" corresponds to many different types in OSM, as tags

a) A building that is named directly

```
<way id="21572801">
  <tag k="building" v="church"/>
  <tag k="name"
    v="Engelbrektskyrkan"/> </way>
```

b) A building with an additional node placed inside that has its name associated to it

```
<way id="163966736">
  <tag k="building" v="yes"/> </way>
<node id="1340902455"
  lat="59.345" lon="18.067">
  <tag k="name" v="Tyskaskolan"/>
</node>
```

Figure 3: Ambiguity in representation: How entities are name-tagged.

specify the size and function of the street, e.g. residential or cycleway. Likewise, describers can use a variety of expressions to refer to the same type, e.g. they could also refer to a street as "a road". Therefore, we need an appropriate mapping to infer the possible matches.

Besides geographic knowledge, more general knowledge about certain objects can be useful to infer their properties even when they are not explicitly mapped. Consider a user that interacts with a navigation system saying "I am following the footpath" but the matching OSM entity is tagged as a bicycle path. In this kind of application, it is useful to assume that bicycle paths can usually be accessed by pedestrians and the RE can be resolved to it.

5 Mismatches Between Map Representation and Speakers' Conceptualization

As mentioned before, OpenStreetMap contains a number of inconsistencies in how entities are tagged. This implies that several strategies can be needed to resolve the same kind of reference. Figure 3 shows the case of names for buildings. A building of any kind (an OSM way), can be tagged with a name directly (3a), or there can be an additional node placed inside the building, that is tagged with the name (3b). In the map representation, there is no direct link between the way and the named node. This connection has to be inferred by computing whether the node's position is inside the building.

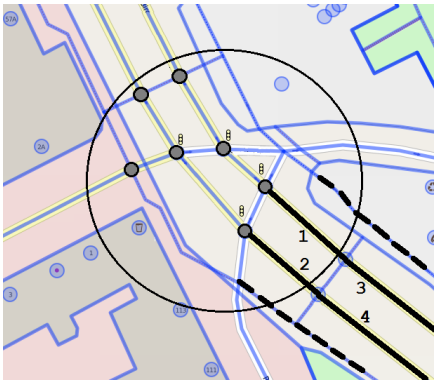

Figure 4: Granularity in OpenStreetMap: an intersection consisting of many street segments and nodes where they meet. The highlighted nodes inside the circle are all part of "an intersection". The highlighted street segments (1-4) belong to the same named street, that is also mapped with a footway and a cycleway running next to it (indicated by the discontinuous lines)

Another problematic case is the granularity with which objects are mapped. Figure 4 shows a major intersection, containing many street segments and nodes where they meet. In the description "I am approaching a junction" it is not at once clear which entities an algorithm should pick.

Grouping larger objects together, such as street segments or buildings that form a unit such as a university campus, is challenging as well. At first sight, this problem could be solved on the basis of the entities' names. Consider however the mapping of large roads, where sometimes the pedestrian walkway is mapped separately, parallely to the road. These pedestrian ways frequently do not contain a name tag and can thus not be associated to the road easily. Additionally, ways can (and often do) consist of several segments, each an own entity in OSM. In Figure 4, each thick black line corresponds to the segment of a street that stretches further in both directions, and has a pedestrian way mapped next to it. Speakers will often refer to the whole structure as "the street" and we need to decide which entities this should correspond to.

6 Resolving References

Keeping the above difficulties in mind, the task is now to map from a referring expression to the user's intended referent, which may be one or more OSM entities.

Referring Expression	OSM tag/value
"road", "street"	highway={tertiary, secondary, primary, residential, pedestrian}
"path", "footpath"	highway={footway, cycleway}
"cycle path", "bike lane"	highway=cycleway
"trees"	natural=tree_row, leisure=park
"traffic lights"	highway=traffic_signals, crossing=traffic_signals
"bus station"	highway=bus_stop
"stairs", "staircase"	highway=steps
"parking lot", "parking space"	amenity=parking
"arches", "archway"	tunnel=yes

Table 1: A set of mappings from referring expressions to features of the OSM entities that the expressions refer to

We can distinguish the following cases:

1. There are zero referents in the database (i.e. the intended referent is not in the database).
2. The intended referent is a unique OSM entity with a single OSM identifier.
3. The intended referent is a unique set of referents in the database ("the two bus stops").
4. A referent can be chosen from a set of interchangeable (equally good) entities in the database.

In the latter case, we either need to devise a mechanism to group the entities together, or we can pick one of them, as the following two examples show:

- "the intersection" can refer to a group of several nodes where street segments meet to form what the speaker perceives as a unit. In this case, we do not want to pick out one of the nodes, but treat them as a unit so that they reflect the extension of the intersection as in expressions like "Cross the intersection".
- "an entrance to the tunnelbana station" can be the building that is the actual entrance, or the node inside it, that is tagged as subway_entrance.

6.1 OSM Features for Resolving References

We have matched all 398 REs in our data with the OSM entity or entities that we judge correspond to the user's intended referent. In 354 cases (89%) the intended referent is present in the database. For all of these 354 REs and correponding referents, we computed the following binary features:

osmName True if the name used in the RE matches the OSM name. We count only exact matches, i.e. the OSM tag name has to exactly match the string in the RE. This serves to give a first overview of how many expressions can be resolved purely by checking the name.

osmName+ True if the name used in the RE matches the OSM name, with some simple normalization using a robust parser. Here, we are applying simple rewriting rules (the RE "the Serbian embassy" is mapped to name="Embassy of the Republic of Serbia") as well as translations of type specifications, such as mapping "Engelbrekt's church" to name="Engelbrektskyrkan"). Note that we are only considering a small part of OSM and additional rules may be needed for cases that we did not come across in this dataset.

osmType True if the type used in the RE (e.g. "restaurant") exactly matches the OSM type. In OSM, types are represented as tags, either as the tag name (building=yes), or as its value (tourism=artwork).

osmType+ True if the type used in the RE matches the OSM type modulo the taxonomy in Table 1 (i.e. the RE "street" matches all the OSM types tertiary, secondary, etc., and "car park" matches entities that are tagged as amenity=parking etc.)

closest True if the entity is the closest of its type to the speaker.

direction True if the entity is located in the speaker's walking direction. For this feature, we are using the previous location of the speaker to define her current bearing. An entity is in her walking direction if it is located within an angle from -90 to 90 degrees (cf. Figure 2).

	Describer		
	A	B	C
# ref. expr.	150	122	126
# in OSM*	134	109	111
name references	14	3	17
`osmName`	3	2	8
`osmName+`	13	3	16
type references	128	106	111
`osmType`	29	45	49
`osmType+`	117	100	102
`closest`	101	75	84
`direction`	130	106	109
`visibility`	125	106	105

Table 2: Counts of referring expressions that can be linked to OSM features as described in Section 6.1 *the OSM data was downloaded in June 2013

visibility True if the entity is visible from where the speaker is. This feature reflects actual visibility, i.e. as judged by the annotators from their knowledge of the environment. An entity can also be visible if it is behind the speaker.

6.2 Results

Table 2 shows the result of the annotation. We can see that the majority of REs contain a type, but that they exactly match the type names and tags in OSM in less than half of the cases. For describer C, all REs contain a type identifier (111), but only 49 of them can be related to their referent without further processing. Applying the mappings shown in Table 1 can improve the matching to more than twice the amount. This is the case for the describers A and B as well.

Very few names were used. However, recall that the describers were asked not to use street names. Consequently, the amount of names might have been higher if they had been allowed to do so.

Furthermore, the table shows that most of the objects are in front of and visible for the speaker (e.g. 97% and 93% for describer A, respectively). In fewer cases (ca. 69–75%), the object was the closest of its type. Note that these three features depend on the position of the speaker and that the GPS signal on which we base these features, varies in accuracy.

The counts in Table 2 show that we can map the type and name of an entity as they are used in the RE with the annotation used in OSM, for a large number of cases. This will limit the number of possible referents, but not suffice to find the actual referent.

Feature combination	Referents found
osmType, osmName, closest	.27
osmType, closest	.30
osmType+, osmName+, closest	.67
osmType+, closest	.68
osmType+, closest, visibility	.65
osmType+, closest, visibility, osmName+	.65
osmType+, closest, visibility, osmName+, direction	.63

Table 3: Applying different combinations of features to resolve references.

In Table 3, we are considering different subsets of the features. We are considering the 354 REs of all three speakers, for which we know that the referent is in the database. The combination of features that covers most mappings uses only the type feature along with the taxonomy in Table 1 (osmType+), combined with the distance information (closest).

Based on these counts, a baseline method can proceed in the following way to find a referent:

1. Compute the set of geographic entities in the vicinity of the speaker's position.
2. From this set, compute the set of possible referents by determining how the entitites are related to one another. At this step, potential referents can be added for entities that make up a unit, e.g. nodes of an intersection as depicted in Figure 4.
3. Filter away entities that do not match the RE in name or type.
4. Pick the closest of the remaining entities.

Note that visibility can be handled in different ways: When computing the initial set of available referents, or at a later point. The counts in Table 3 reflect a lower number of matches when including information about visibility. This may be because of inaccuracies in the GPS signal, or simply an artefact of the small dataset.

7 Discussion and Future Work

The ultimate aim of this work is to develop a robust reference resolution method that can be

incorporated into our pedestrian navigation system (Boye et al., 2014). Therefore, it is important to point out that the above results were all obtained using data where users described the way as they were walking, and consequently it was natural to resolve a spatial reference to a matching entity closest to the user's position. However, there are situations where users would refer to entities and places that are possibly far away (e.g. "How do I get to X street?"). Therefore any realistic spatial reference algorithm must take the user's dialogue act into account: For instance, if the user is making a request ("Give me directions to X"), proximity to X should not be given much weight.

Furthermore, in this paper we have only considered how many of the intended referents we can find, but it is also important to identify the references that have no referent in the database, as to avoid false positives. Such a procedure needs to make an assumption about the coverage of OSM in a particular area as well.[6]

As discussed before, it is often far from obvious what the intended referent is. In particular this is true in situations where the user conceptualizes her surroundings differently from how the database is organized (as in Figure 4). A possibility would be to add an extra layer on top of OpenStreetMap, in which nodes are grouped into super-concepts like "intersection", "roundabout", etc. Such super-concepts could be formed on the basis of actual data, like the verbal route descriptions we are using in this study. This would have the advantage of resolving references to entities that more closely correspond to the user's mental map, but the disadvantage of requiring extra computation.

Additional processing is also required when the reference resolution is to be carried out in other languages than English. In our features, we exploited the fact that OSM tags and values are in English and therefore match natural language expressions in some cases. Further linguistic processing and algorithms that map OSM concepts to language resources such as WordNet, like *Voc2WordNet* (Ballatore et al., 2014), may be a useful resource to bridge the gap between commonly used terms and map concepts.

[6] A visualization of the OSM coverage can be found at `https://www.mapbox.com/osm-data-report/`

Acknowledgment

The authors were supported by Swedish national grant VR 2013-4854 "Personalized spatially-aware dialogue systems".

References

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. 2004. Web-a-where: Geotagging Web Content. In *Proc. of SIGIR*, pages 273–280.

A. Ballatore, M. Bertolotto, and D. C. Wilson. 2014. Linking geographic vocabularies through WordNet. *Annals of GIS*, 20(2):73–84.

N. Blaylock. 2011. Semantic Annotation of Street-level Geospatial Entities. In *Proc. of the IEEE ICSC Workshop on Semantic Annotation for Computational Linguistic Resources*.

J. Boye, M. Fredriksson, J. Götze, J. Gustafson, and J. Königsmann. 2014. Walk This Way: Spatial Grounding for City Exploration. In *Natural Interaction with Robots, Knowbots and Smartphones*, pages 59–67. Springer New York.

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

J. Götze and J. Boye. 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.

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

B. Martins, M. J. Silva, S. Freitas, and A. P. Afonso. 2006. Handling Locations in Search Engine Queries. In *Workshop on Geographical Information Retrieval, SIGIR*.

C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox. 2014. Learning from Unscripted Deictic Gesture and Language for Human-Robot Interactions. In *Proc. of the 28th National Conference on Artificial Intelligence*.

R. Mitkov, 2010. *Computational Linguistics and Natural Language Handbook*, chapter Discourse processing, pages 599–629. Blackwell Publishers.

B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.-C. Forslund, and C. Best. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *Proc. of LREC-2006*.

A. Tom and M. Denis. 2004. Language and spatial cognition: comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology*, 18(9):1213–1230.