Towards Machine-based Matching of Addresses Expressed in Natural Languages

Farid Karimipour*, Ali Javidaneh*, Andrew U. Frank**

* Department of Surveying and Geomatics Engineering, College of Engineering, University of Tehran, Iran
** Research Group Geoinformation, Department of Geodesy and Geoinformation, Vienna University of Technology, Austria

Abstract. Address matching is frequently used in everyday life. It is an essential prerequisite for many of the functionalities provided by location-based services (e.g. car navigation). The procedure is simply parsing an address expressed in a pre-defined standard format to its components, and then matching these components with their corresponding features on the map. If such standards are absent however, the parsing and, consequently, the matching usually fail; thus human intervention is needed. As one of such cases, this paper presents the initial results of an ongoing research on developing a machine-based matching for addresses expressed in natural languages. As the first step, we show how such addresses can be parsed through formal expression of their combination rules. The implementation result for a case study is presented.

Keywords. Address matching, Addressing standards, Natural languages, Parsing, Location-based services, BNF

1. Introduction

Address matching (also called *geocoding*) is an applied spatial analysis which is frequently used in everyday life. Almost all desktop (e.g. ArcGIS) and web-based (e.g. Googlemaps) GIS environments are equipped with a module to match the addresses expressed in pre-defined standard formats on the map. It is an essential prerequisite for many of the functionalities provided by location-based services (e.g. car navigation). ERSI defines address matching as "a process that compares an address or a table of addresses to the address attributes of a reference dataset to determine wheth-

Published in "Proceedings of the 11th International Symposium on Location-Based Services", edited by Georg Gartner and Haosheng Huang, LBS 2014, 26–28 November 2014, Vienna, Austria.

er a particular address falls within an address range associated with a feature in the reference dataset. If an address falls within a feature's address range, it is considered a match and a location can be returned" (ESRI's GIS Dictionary on Address Matching). Several methods have been proposed for address matching which assume a standard format for the components of the address, and propose solutions for matching the known address components to map components (O'Reagan & United States. Bureau of the Census. Statistical Research 1987; Yu 1996; Yang et al. 2004; Goldberg et al. 2007; Goldberg 2011; Eckman & English 2012; Qin et al. 2013).

Address matching is composed of four main steps: (1) Text parsing, (2) Standardization, (3) Correction, and (4) Matching (Yang et al. 2004). Although address matching may be considered a straightforward, well-studied issue, there are still many questions to be answered. Surfing the web, you encounter practical questions such as: "the same address may be referred to in multiple ways: 110 Test St, 110 Test St., 110 Test Street, etc.", where different notations are used for "street". A more complex situation is: "the addresses could be written in all different ways: 1345 135th St NE, 1345 NE 135th St, etc.", in which the order of components may change. Such problems may be partially solved by designing a comprehensive parser (step 1) that captures all of the above addressing formats. The parsed address is then standardized (step 2) and corrected (steps 3) before performing the final matching (step 4). However, unmatched addresses may still remain (30% on average), which then require user intervention (Yang et al. 2004).

In both of the above examples the components of the addresses are fixed, and only the symbols (the first example) or the order (the second example) differ. The situation is much more complicated when there is no addressing standard present whatsoever, and addresses are expressed in natural languages (hereafter called *textual addresses*). An interesting example is Iran, where people express addresses as a sequence of spatial elements (e.g. streets, squares, landmarks, etc.), starting from a known element. For example, in *Figure 1* the address of point **A** based on route #1 is "Tehran, Shariati Ave., Gholhak, Pabarja St., Ayeneh Blvd., West corner of Gol-e-yakh Alley, No. 2, Unit 9". Even worse, the same place could also be referred to in completely different ways because different starting points or spatial elements may be used by another person. For example, based on route #2 in *Figure 1*, point **A** is referred to as "Tehran, Daroos, Shahrzad Blvd., Pabarja St., Ayeneh Blvd., West corner of Gol-e-yakh Alley, No. 2, Unit 9". Note that in the first address, "Gholhak" is an old famous name for the point shown on the map; and in the second address, "Daroos" is an old name for this area, which is not even shown on the map!

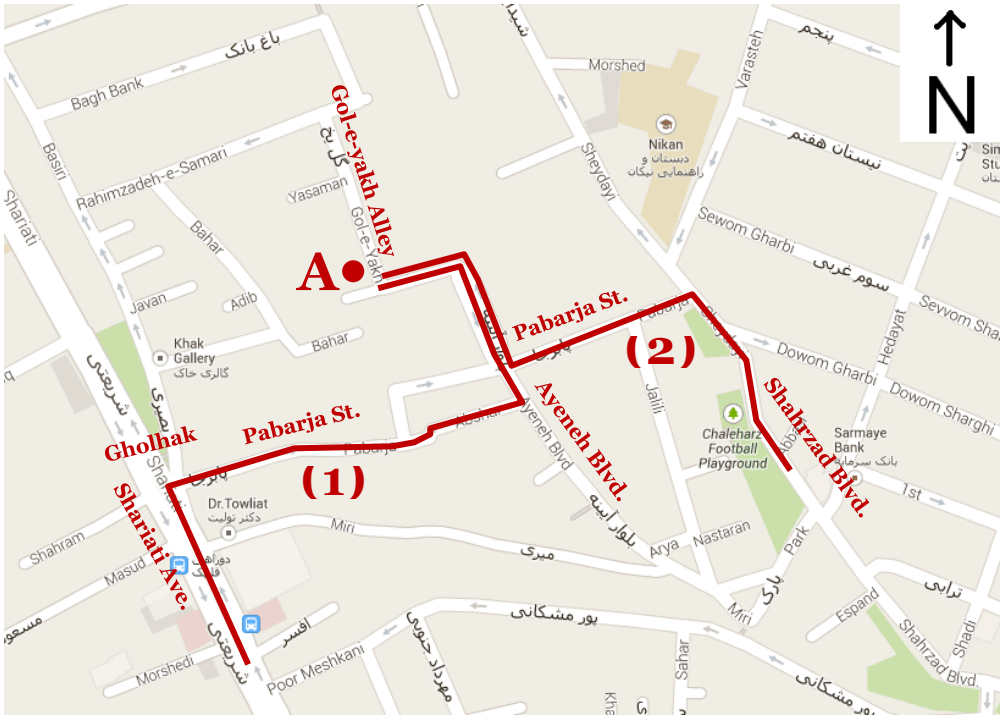

Figure 1. Point **A** is referred to in two different ways using different stating points and spatial elements.

Although this method of addressing may seem very unpleasant at first, it is very efficient, because:

- It not only specifies the destination, but it also tells how to reach it. In other words, you do not need any map, navigation system, etc. to find the destination. Instead, you can reach the known starting point and then look for the next components, step-by-step. Compare this with "Bräuhausgasse 64/7, 1050 Vienna", which only tells you that your destination is located in the 5th district of Vienna, but you will need to look all over the 5th district on the map for Bräuhausgasse, or search for it in your navigation system.
- This way, you will inevitably be exposed to the environment and its spatial elements, which helps you in building up your cognitive map. Again, compare it with having a navigation system that tells you how to reach Bräuhausgasse. In this case, you rely on the navigation system and do not necessarily have any connection to the environment.

Nevertheless, even if, as we claim, this method of addressing is efficient, it prevents many of the location-based services that require address matching to be (efficiently) used. A simple practical example is car navigation sys-

tems: even if you know your destination, you have to find it and point to it on the map through zooming and panning; very difficult to do for unknown destinations!

This paper presents the initial results of an ongoing research on developing a machine-based matching for addresses expressed in natural languages. We believe that text parsing is the most critical step in this regards. Such non-standard addresses are expressed in natural languages (Chomsky 1956; Ginsburg 1975; Bolc & Carbonell 1987; Allen 1995; Chomsky 2002; Gómez-Rodríguez 2010; Biemann 2012; Pustejovsky & Stubbs 2013) and based on the user's spatial cognition (Lynch 1960; Coventry et al. 2009; Frank 2010; Goodchild 2011; Hirtle 2011; Karimipour & Niroo 2013; Khazravi & Karimipour 2012). Therefore, the elements of spatial cognition, as well as linguistic structures must be considered in order to efficiently parse such textual addresses. The attempts to deploy the machine-based natural language processing (NLP) in address-matching has reached limited successes. The main reason is the complexity of natural language structures, which caused an address to be unmatched even if there is a partial error in its parsing.

Language is conceptually connected to human cognition. Twaroch & Frank (2003) combined language and space for better understanding of spatial cognition and urban environments. Shusterman et al. (2011) studied the role of language in development of spatial cognition. There are other studies on connections between space and language (Talmy 1983; Bloom 1999).

In this paper, we propose a simple formal language to parse the addresses expressed in natural languages. Section 2 describes how formal languages may be used to formally express the combination rules of natural languages. This results in a simple formal language to parse the textual addresses expressed in natural languages, which is presented and tested for a case study in Section 3. Lastly, Section 4 concludes the paper and introduces future research directions.

2. Formal Expression of Combination Rules of Natural Languages

A *formal language* consists of a set of *symbols* (equivalent to words in natural languages) and a set of rules for their combination, called *syntax* (roughly equivalent to grammar in natural languages). A valid combination of symbols is called a *well-formed construction*, which can be parsed and interpreted using the combination rules.

Although natural languages do not obey such strict combination rules, still a level of regularity can be considered, and thus the parsing methods used for formal languages may be applied (Chomsky 1980). In particular, an address expressed in a natural language may be automatically parsed to its components by defining a set of combination rules, i.e. grammar.

Grammar is a set of rules to combine the symbols (vocabularies) of a natural language (Frank 2006). The symbols are classified into (*Figure 2*):

- A set of *terminal symbols*, which cannot be further simplified. Terminal symbols may be *constants* or *variables*.
- A set of *non-terminal symbols*, which must be eventually simplified to the terminal symbols by repeated application of combination rules.
- A special non-terminal symbol *S*, which is called the *start symbol*.

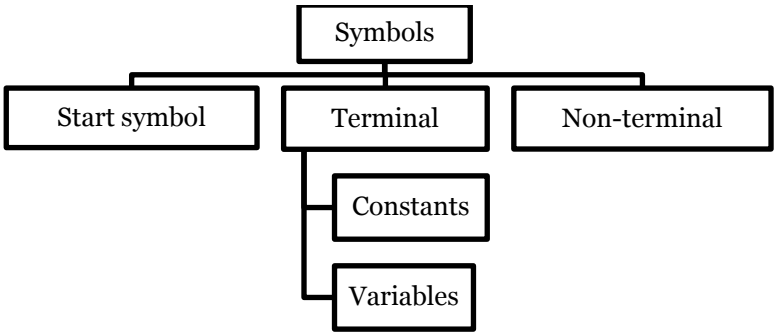

Figure 2. Classification of symbols of a language.

The combination rules are usually expressed in Backus-Naur Form (BNF), which is a formal meta-language to define other languages (Frank 2006). *Table 1* illustrates the main symbols used in the BNF.

:=	is replaced by or produces
|	or
[]	optional
{}	any number
()	grouping
“”	enclose terminal symbols

Table 1. The main symbols of the BNF.

For example, the grammar of a simple natural language may be defined using the BNF as follows:

1. S:=NP VP
2. NP:= M1 N M2
3. VP:= V NP
4. M1:= article | quantifier | N | demonstration | superlation
5. M2:= appositive | relative clause

where NP, VP, M1 and M2 stands respectively for “Noun Phrase”, “Verb Phrase”, “Modifier type 1” and “Modifier type 2”. Figure 3 illustrates how a well-formed construction in this language is parsed using the specified grammar.

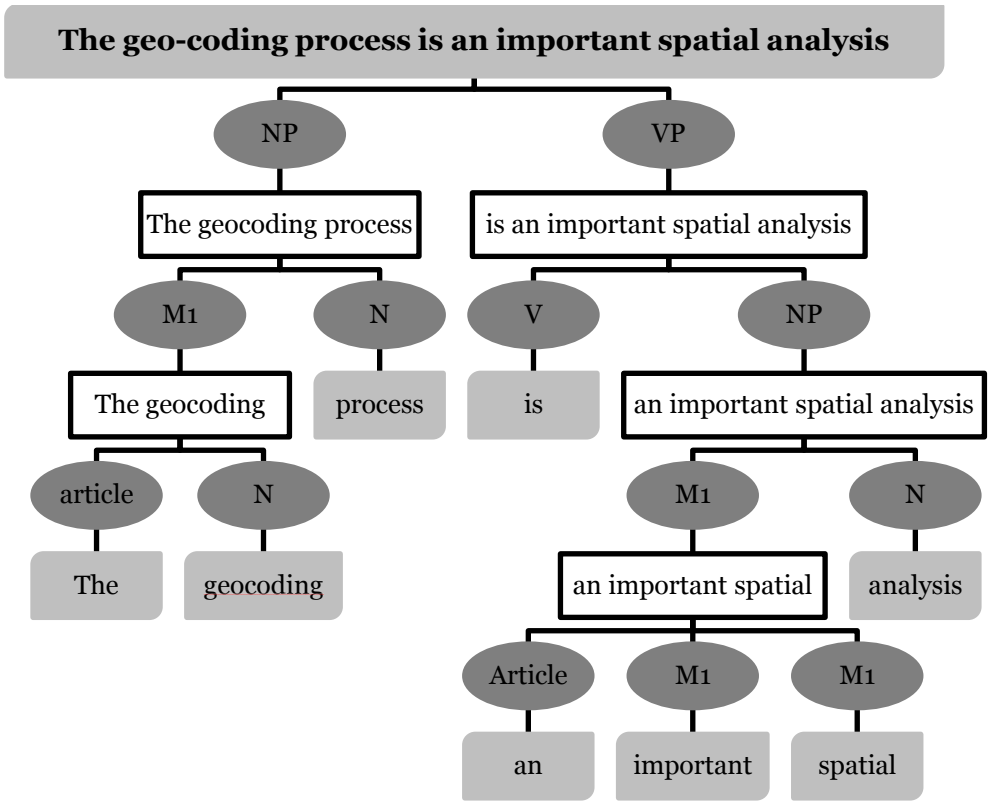

Figure 3. Parsing a well-formed construction using the specified grammar.

3. A Language to Parse Textual Addresses: A Small Example

As discussed, in order to parse textual addresses a formal language including valid grammar and vocabularies must be defined. As textual addresses are expressed in natural languages, the grammar must be as flexible as possible in order to optimally capture irregularities.

A textual address may consist of two *spatial groups (SG)* of terminal symbols:

1. *Geo-names (GN)*
   1.1. *Constant geo-names (cGN)*: avenue, street, alley, etc.
   1.2. *Variable geo-names (vGN)*: names of the constant geo-names
2. *Spatial relations (SR)*: after, before, etc.

Therefore, the combination rules of such a language may be defined as:

1. S := {SG, #41}
2. SG := [SR] GN
3. GN := cGN vGN | vGN cGN
4. cGN := “avenue” | “ave.” | “street” | “st.” | “alley” | “number” | “unit”
5. SR := “after” | “before” | “in front of” | “left of” | “right of”

Figure 5 illustrates the parsing of a sample textual address, whose position is shown in *Figure 4*, using the language defined.

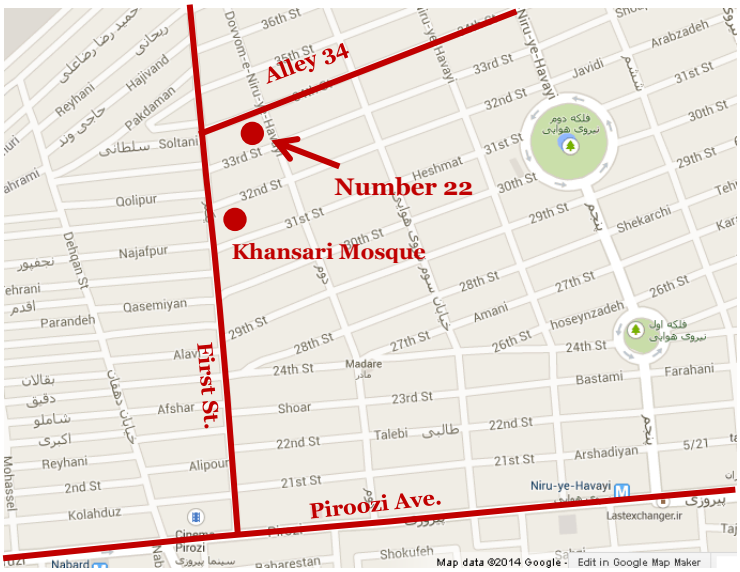

Figure 4. Position of the sample textual address on the map.

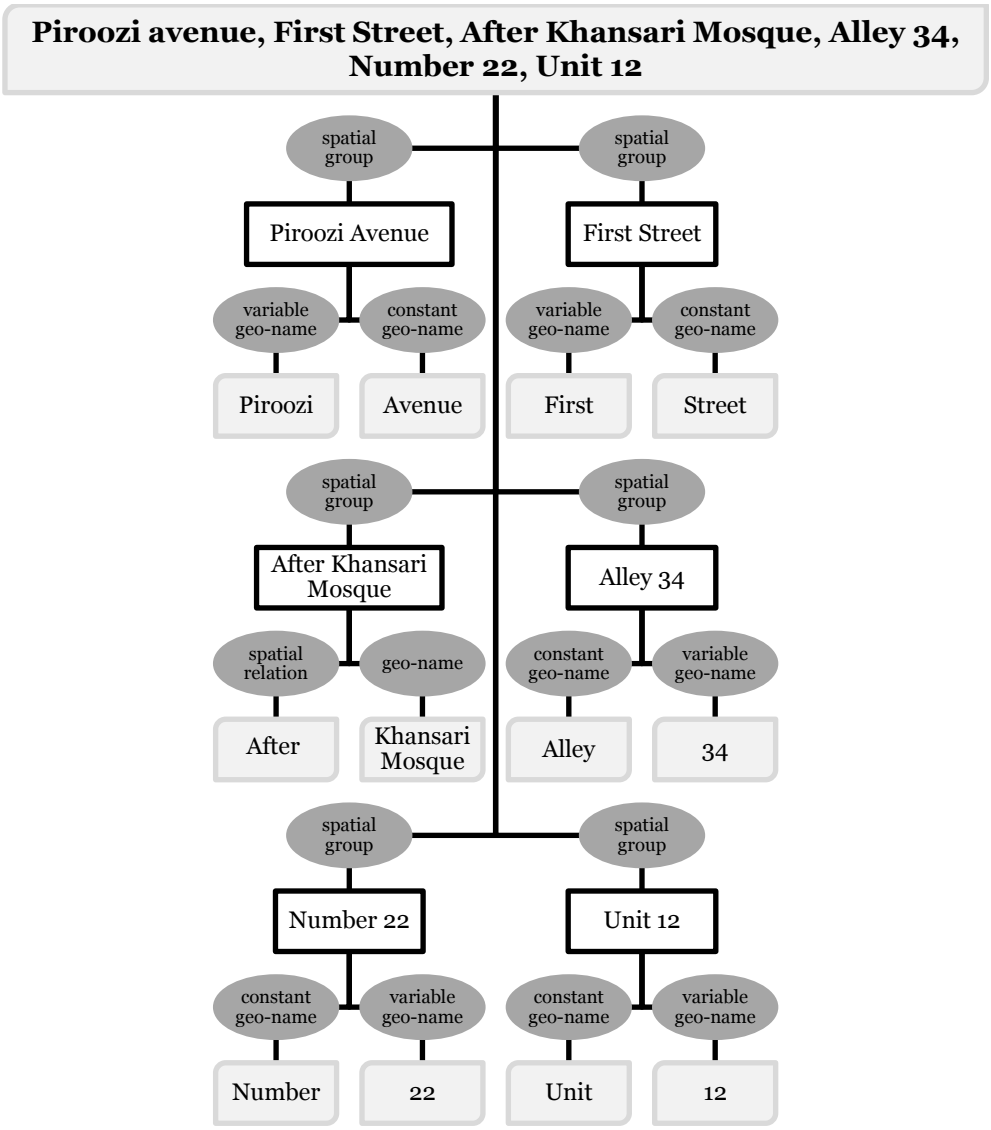

Figure 5. Pasring of a sample textual address using the language defined.

The proposed approach was implemented as a simple software (*Figure 6*), programmed in Microsoft Visual Studio C#.net. The user can introduce the geo-names and spatial relations (left). Then, any textual address that contains those geo-names and spatial relations is parsed to its components (right), which can be used by the next address matching steps.

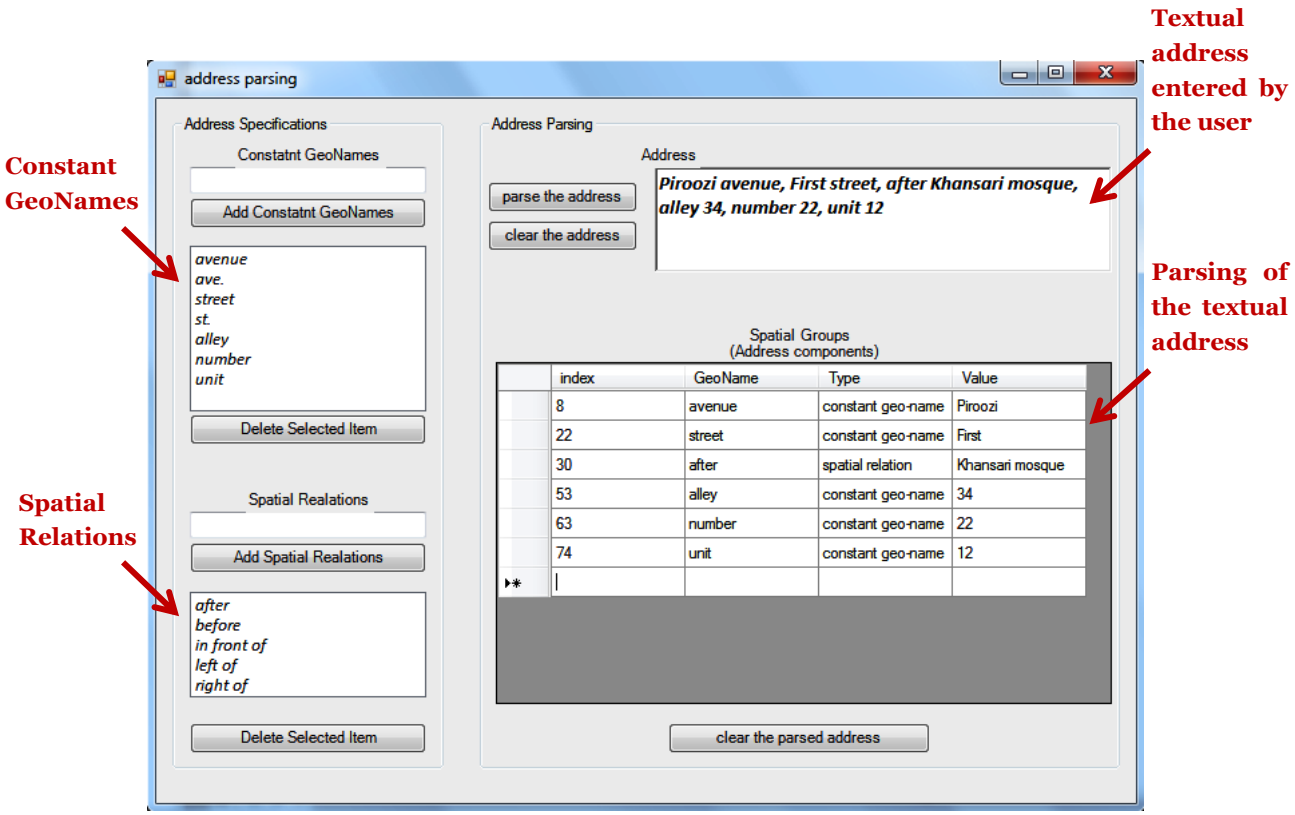

Figure 6. Implementation of the proposed approach of the paper: The textual address (top right) is parsed (bottom right) based on the geo-names (top left) and spatial relations (bottom left) introduced by the user.

4. Conclusions and Future Work

This paper discussed the initial results of an ongoing research on developing machine-based matching for addresses expressed in natural languages. As the first step, we defined a formal language - including symbols and combination rules - to parse such textual addresses. The implementation result seems promising, as we could parse a sample address to its components. However, it is still a long way to practically developing the desired address matching. The most problematic issue is that natural languages do not completely obey the rules provided by formal languages. We are currently working on enriching the proposed language with more symbols and combination rules in order to capture as many irregularities as possible.

References

Allen J (1995) Natural Language Understanding. Benjamin/Cummings Publishing Company. Redwood City, California

Biemann C (2012) Structure Discovery in Natural Language. Springer

Bloom P (1999) Language and Space. MIT press

Bolc L, Carbonell JG (1987) Natural Language Parsing Systems. Springer-Verlag. Berlin; New York

Chomsky N (1956) Three Models for the Description of Language. Information Theory, IRE Transactions on 2 (3):113-124

Chomsky N (1980) Rules and Representations. The Behavioral and Brain Sciences 3:1-61

Chomsky N (2002) Syntactic Structures. Walter de Gruyter

Coventry KR, Tenbrink T, Bateman JA (2009) Spatial Language and Dialogue. Oxford University Press. Oxford; New York

Eckman S, English N (2012) Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates. Field Methods 24 (4):399-408.

ESRI's GIS Dictionary on Address Matching. http://support.esri.com/en/knowledgebase/GISDictionary/term/address%20matching. Accessed 3 May 2014

Frank AU (2006) Practical Geometry: The Mathematics for GIS, Course Lecture

Frank AU (2010) How Do People Think About Space. Position paper for Dagstuhl seminar 10131

Ginsburg S (1975) Algebraic and Automata-Theoretic Properties of Formal Languages. Elsevier Science Inc

Goldberg DW (2011) Advances in Geocoding Research and Practice. Transactions in GIS 15 (6):727-733

Goldberg DW, Wilson JP, Knoblock CA (2007) From Text to Geographic Coordinates: The Current State of Geocoding. URISA Journal 19 (1):33-47

Gómez-Rodríguez C (2010) Parsing Schemata for Practical Text Analysis. Imperial College Press

Goodchild MF (2011) Spatial Thinking and the Gis User Interface. Procedia - Social and Behavioral Sciences 21:3-9

Hirtle SC (2011) Geographical Design Spatial Cognition and Geographical Information Science. Morgan & Claypool

Karimipour F, Niroo A Agent-Based Modelling of Cognitive Navigation: How Spatial Knowledge Is Constructed and Used. In: Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on, 2-5 Dec. 2013 2013. pp 17-22

Khazravi A, Karimipour F (2012) Cognitive Readability Enhancing of Cartographic Maps for Pedestrian Navigation. International Journal of Brain and Cognitive Sciences 1 (3):11-17

Lynch K (1960) The Image of the City, vol 11. the MIT Press

O'Reagan RT, United States. Bureau of the Census. Statistical Research D (1987) Geocoding Theory and Practice at the Bureau of the Census. Bureau of the Census. Washington, D.C.

Pustejovsky J, Stubbs A (2013) Natural Language Annotation for Machine Learning. O'Reilly

Qin X, Parker S, Liu Y, Graettinger AJ, Forde S (2013) Intelligent Geocoding System to Locate Traffic Crashes. Accident Analysis & Prevention 50 (0):1034-1041

Shusterman A, Ah Lee S, Spelke ES (2011) Cognitive Effects of Language on Human Navigation. Cognition 120 (2):186-201

Talmy L (1983) How Language Structures Space. Springer

Twaroch F, Frank AU (2003) Improving the Understanding of Space through Language. Paper presented at the COSIT

Yang DH, Bilaver LM, Hayes O, Goerge R (2004) Improving Geocoding Practices: Evaluation of Geocoding Tools. Journal of medical systems 28 (4):361-370

Yu L (1996) Development and Evaluation of a Framework for Assessing the Efficiency and Accuracy of Street Address Geocoding Strategies. State University of New York at Albany