Inter-Annotator Agreement on a Linguistic Ontology for Spatial Language
A Case Study for GUM-Space

Joana Hois

Research Center on Spatial Cognition (SFB/TR 8)
University of Bremen, Germany
P.O. Box 330 440, 28334 Bremen
joana@informatik.uni-bremen.de

Abstract

In this paper, we present a case study for measuring inter-annotator agreement on a linguistic ontology for spatial language, namely the spatial extension of the Generalized Upper Model. This linguistic ontology specifies semantic categories, and it is used in dialogue systems for natural language of space in the context of human-computer interaction and spatial assistance systems. Its core representation for spatial language distinguishes how sentences can be structured and categorized into units that contribute certain meanings to the expression. This representation is here evaluated in terms of inter-annotator agreement: four uninformed annotators were instructed by a manual how to annotate sentences with the linguistic ontology. They have been assigned to annotate 200 sentences with varying length and complexity. Their resulting agreements are calculated together with our own 'expert annotation' of the same sentences. We show that linguistic ontologies can be evaluated with respect to inter-annotator agreement, and we present encouraging results of calculating agreements for the spatial extension of the Generalized Upper Model.

1. Introduction

In the field of natural language processing (NLP), ontologies can be used as the central component of representing knowledge, in particular they can either be used to identify domain knowledge or semantic categories (Chandrasekaran et al., 1999). The combination of both plays an essential role in understanding discourse of a given domain. In this paper, we present an evaluation of a linguistic ontology that specifies such semantic categories of natural language, namely the Generalized Upper Model (Bateman et al., 1995). It groups together distinguishable meanings that language itself constructs by providing a categorization into ontological classes and relations accordingly.

The Generalized Upper Model (GUM) provides semantic descriptions of parts of natural language sentences, and it supports the relation to contextualizations of these sentences. In particular, the GUM descriptions address just the degrees of underspecification that a linguistic utterance itself leaves open, while they define precisely what can be extracted from an utterance's semantics (Bateman et al., 1995). GUM has been further extended to specify those categories that are relevant for natural language of space (Bateman et al., forthcoming). It builds on related work in natural language research, such as (Talmy, 2006), (Levinson, 2003) and (Halliday and Matthiessen, 1999), as well as on empirical analysis of natural language corpora, in order to construct appropriately motivated semantic types.

In general, the quality of an ontology can be measured with regard to different types of criteria: (1) *lexical* criteria evaluate appropriateness and intelligibility of used terms in the ontology; (2) *structural* criteria analyze metrical aspects of an ontology in terms of complexity, reasoning, and graph structure properties; (3) *representational* criteria evaluate whether an ontology adequately formalizes its intended domain; (4) *application* criteria test whether an ontology supports certain applications in its intended ways; (5) *usability* criteria evaluate availability and re-usability of an ontology; and (6) *philosophical* criteria analyze formal ontological ideas of an ontology (cf. (Brank et al., 2005)). In particular, NLP techniques can further evaluate linguistic aspects of an ontology by measuring its precision and accuracy (Obrst et al., 2007).

As GUM's spatial extension primarily aims at categorizing spatial language and thus enabling dialogue systems to understand spatial language (cf. (Ross, 2008) for an application-oriented analysis), we will here focus on evaluating representational criteria by measuring inter-annotator agreements. The next section introduces the spatial extension of GUM and illustrates how the ontology specifies natural language of space. We subsequently present how we conducted an inter-annotator agreement study for the spatial extension. Finally, we present the results of this study and discuss future work.

2. The Spatial Extension of the Generalized Upper Model

The linguistic ontology GUM-Space[1] is developed formally as an ontological extension (Konev et al., 2009) of GUM by refining those components that are necessary to specify detailed information in spatial language utterances, primarily for English and German (Bateman et al., 2007). GUM itself is intended to be used within natural language dialogue systems by providing a specification of the semantics of language. GUM-Space has particularly been applied to a natural language dialogue system for spatial assistance (Ross et al., 2005). Furthermore, given the formal semantics that GUM-Space provides, it can be used for interpreting spatial expressions in a situational context, for instance, in relation to formal representations of spatial scenes (Hois and Kutz, 2008b; Hois and Kutz, 2008a).

[1] GUM-Space is accessible online at http://www.ontospace.uni-bremen.de/ontology/stable/GUM-3-space.owl

GUM-Space can be used for specifying spatially-related expressions in natural language, and we briefly illustrate the way it categorizes spatial language. In principle, sentences can be analyzed with respect to their contributing meaning to the spatial meaning by the following aspects:

- spatial location, position, positioning, orientation, or movement of entities
- spatial positions, locations, places described by directions, paths, orientations, starting points, end points, or intermediate points of a route
- spatial relations between entities
- spatial modification of any of the previous spatial information (e.g., a specific distance, an angle, a perspective)
- collections, combinations, or connections of all such spatial information

Similar meanings or variations of this kind of information are categorized into the same groups in GUM-Space, which specifies them as particular roles or categories based on their ontological representation. The sentence "The chair is next to the table.", for instance, is categorized as:

SpatialLocating *sl1*:
 locatum: *chair*
 processInConfiguration: *is*
 placement: GeneralizedLocation *gl1*:
  hasSpatialModality: RelativeNonProjectionAxial *rnpa1*
  relatum: *table*

Here, a configuration SpatialLocating (an ontological class) specifies that the spatially-relevant information in the sentence refers to a static spatial position of an entity (the locatum, an ontological relation), which is the "chair" (an ontological class). This entity is related to a certain placement that consists of a spatial relation (hasSpatialModality) and the related reference object (relatum), which is the "table".[2] GUM-Space provides about 70 different categories for spatial relations (the SpatialModality) that define how entities can be located in space with respect to certain linguistic and environmental constraints. One of them is RelativeNonProjectionAxial, which reflects the relative position between two entities based on their axial alignment, expressed, for instance, in "A is besides B" or "C is next to D".

In a similar way, information on motions, orientations, routes, directions, perspectives, and modifications can be specified. The example sentence "From there carry on to the end of the road that you are on.", for instance, is specified in GUM-Space as follows:

NonAffectingDirectedMotion *nadm1*:
 actor: *you*
 processInConfiguration: *carry on*
 route: GeneralizedRoute *gr1*:
  destination: *GeneralizedLocation gl1*:
   hasSpatialModality: GeneralDirectionalNearing *gdn1*
   relatum: *end of the road*
   spatialPerspective: *from there*

AND SpatialLocating *sl1*:
 locatum: *end of the road*
 processInConfiguration: *undefined*
 placement: GeneralizedLocation *gl2*:
  hasSpatialModality: Peripheral *p1*
  relatum: *road*

AND SpatialLocating *sl2*:
 locatum: *you*
 processInConfiguration: *are*
 placement: GeneralizedLocation *gl3*:
  hasSpatialModality: Support *s1*
  relatum: *road*

This example shows how GUM-Space specifies motion configurations by identifying information about the actor, the type of motion, and the different aspects of a route. The configuration NonAffectingDirectedMotion specifies that the motion configuration is directed, i.e., a route segment is part of the sentence, which in this case is given by the destination "to the end of the road". Route segments are specified within the route category (GeneralizedRoute), that can specify sources, pathPlacements, pathIndications, and destinations. The example also shows that modifications of spatial information is specified as parts of the GeneralizedLocation: The perspective "from there" in the sentence is specified by one type of such modifications. Moreover, the example indicates how GUM-Space breaks down the different spatial units in the sentence into different types of configurations. "From there carry on to the end of the road that you are on." consists of the three different sub-units (1) "from there carry on to the end (of the road)", (2) "the end of the road", and (3) "you are on the road". The latter two are specified as a SpatialLocating configuration similar to the static example above, while the first is specified as a dynamic configuration.

As spatial language sentences can thus be specified by such ontological representations, GUM-Space can be used as an *annotation schema* for natural language of space. Hence, it can be analyzed whether GUM-Space specifications of spatial language are scalable, reliable, and comprehensible, which can be calculated on the basis of *inter-annotator agreement*. If different annotators annotate the same sentences equally, it can be inferred that GUM-Space provides a clear, adequate, and comprehensible distinction between categories and relations. Inter-annotator agreement therefore can be used as a quality criterion to evaluate GUM-Space and a method to guarantee that same meanings of sentences are annotated in the same way. Note, however, that inter-annotator agreement does not intend to prove

[2]Names of the terms in GUM-Space are inspired by the terminology in (Levinson, 2003).

C	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Clean Copy of Original Transcript	Configuration	locatum/ actor	*Term of $H*	process	*Term of $J*	(actee)	*Term of $L*	placement /route/direction	if $N is indicates route information...	hasSpatialModality	*Term of $P*	relatum	*Term of $R*	modification	*Term of $T*
We get an engine from Avon to Bath	AffectingDirectedMotion	actor:	*we*	process:	*get*	actee:	*an engine*	route	source	GeneralDirectionalDistancing	*from*	relatum:	*Avon*	---	-
									destination	GeneralDirectionalNearing	*to*	relatum:	*Bath*	---	-
Send a boxcar with (an) engine back to Avon	AffectingDirectedMotion	actor:	*undefined*	process:	*send back*	actee:	*a boxcar with (an) engine*	route	destination	GeneralDirectionalNearing	*to*	relatum:	*Avon*	---	-
We went via Dansville	NonAffectingDirectedMotion	actor:	*we*	process:	*went*		-	route	pathIndication	PathRepresentingExternal	*via*	relatum:	*Dansville*	---	-
Dansville is three hours from Avon	SpatialLocating	locatum:	*Avon*	process:	*is*		-	placement	---	QuantitativeDistance	*from*	relatum:	*Avon*	quantitativeDistanceExtent	*three hours*

Figure 1: Examples of annotations from the Trains corpus. The upper row shows the different ontology classes and relations to be selected by the annotators, the left column shows the sample sentences, and the main part shows the annotation for these sentences.

that GUM-Space represents human conceptual structures of space or language of space. Instead it proves that GUM-Space's categories can be learned by non-experts and (after a training phase) can be distinguished correctly. It also implies that the categories provide a distinction of groups and that they are not randomly chosen for some linguistic terms. Moreover, and this is the primary aim of GUM-Space, it ensures a coherent representation as an underlying linguistic basis for dialogue systems.

3. Inter-Annotator Agreement Study

Measuring the agreement between two or more annotators, who annotate the same data sample given a certain annotation schema, can prove the consistency and the reliability of the annotation schema (Gwet, 2001). Inter-annotator agreement can analyze the appropriateness, applicability, and comprehension of the categories of the schema. (Lombard et al., 2002) present several criteria and procedure aspects for conducting studies of inter-annotator agreement. Our own inter-annotator agreement study for GUM-Space was mostly guided by this procedure. We evaluate *reproducibility* by comparing annotations between two novice annotators and *accuracy* by comparing the annotations to a golden standard (Krippendorff, 1980). Novice annotators were also independent, i.e., they did not discuss their results (Krippendorff, 1980).

The annotation schema for GUM-Space depends on the specific structure of the linguistic ontology. Annotation schemata most often provide annotations for few categories (or only one) for the *units* (parts of sentences) to be analyzed. Such categories are preferably binary types, an annotator then only needs to check whether a certain condition holds or not. Also, as few categories as possible are supposed to be annotated per unit, in order to keep the annotation schema simple (Lombard et al., 2002). In case of GUM-Space, however, units are annotated as complex structures, as shown above. One sentence is defined by a construction that consists of different relations and categories for each spatially-related unit in the sentence. Moreover there are 70 different types of spatial modalities to distinguish, which are hierarchically structured. Hence, in our study we also investigate whether it is possible at all to apply inter-annotator agreement for GUM-Space, and whether annotators are able to learn complex structures (similar to the annotation shown above for "The chair is next to the table." and "From there carry on to the end of the road that you are on.") and distinguish the different categories for annotating spatial language according to the schema.

For the agreement study, all annotators were provided with a manual for annotating sentences with GUM-Space together with a spreadsheet document file containing a structure of the GUM categories in each column[3]. The annotation task was split into a training phase with 10 sentences, a supervised annotation phase with 2×50 sentences, and an unsupervised annotation sample with 100 sentences. Sentences from the training, supervised, and unsupervised phase did not overlap. Inter-annotator agreement was calculated for the unsupervised annotation samples. All of the sentences were randomly taken from experimental data on spatial language. The English annotation samples are taken from the corpora Trains 93 Dialogues (Heeman and Allen, 1995) and IBL (Instruction Based Learning) Corpus (Lauria et al., 2001), the German annotation samples are taken from the corpora Aibo2 (Fischer, 2007) and Rolland (Shi and Tenbrink, 2009). The study consists of two annotators per language and additionally one 'expert' annotator (the developers of GUM-Space), referred to as the "golden standard" (Gwet, 2001) as the third annotator per language. Annotators were instructed, first, to clean up the sentences, i.e., remove non-spatial information from the sentence and reformulate the sentence if necessary according to the manual instructions, and second, to annotate the sentence according to the GUM-Space specification as given by the

[3]The manual, spreadsheets, and results are available at `http://www.informatik.uni-bremen.de/~joana/gum/gum-iaa.html`

	Category	Percent Agreement (%)	Cohen's Kappa	Cases (No.)	Categories Used (No.)
English	configuration	78.261	0.669	115	8
	spatial role	82.211	0.78	193	12
	modality	72.96	0.694	143	24
	modification	82.211	0.78	15	5
German	configuration	84.804	0.763	136	9
	spatial role	79.588	0.743	178	12
	modality	71.748	0.698	164	36
	modification	71.705	0.657	86	10

Table 1: Average inter-annotator agreements for GUM-Space between three annotators for each language.

spreadsheet document. Examples of annotated sentences are shown in the spreadsheet in Figure 1.

4. Results for GUM-Space

Table 1 shows the calculation results for inter-annotator agreements for GUM-Space between three annotators (two novice annotators and one expert annotator) for English and German respectively.

Calculations for the agreements are split into the four major annotation groups of GUM-Space, namely (1) the configuration of sentences as a whole, (2) the spatial role for placements, directions, or routes, (3) the spatial modality indicating the relative spatial position between the locatum or actor and related entities, and (4) the modification of spatial information, such as angles, perspectives, or distances. This covers the different types of ontological classes and relations introduced in Section 2. The results show agreement above 70% for each annotation group. Given the diversity of up to 36 different categories used, this indicates a promising result that annotators are able to annotate sentences according to the GUM-Space schema and distinguish the different semantic groups. Differences between the results for modifications from English and German depends most likely on the small variety of modification in the (randomly chosen) English sample. The calculation of Cohen's Kappa (Cohen, 1960) is supposed to eliminate agreements between annotators by chance. Average kappa values for all four groups are around 0.7, although the kappa value can go up to more than 0.8 for comparing the annotations between just two annotators. Note, however, that we use the kappa metrics primarily because it is commonly used to measure agreements.[4]

In general, the calculation of agreement was performed in a 'strict' way, i.e., even similar categories with respect to the GUM-Space hierarchy (e.g., LeftProjection and its subcategories LeftProjectionInternal and LeftProjectionExternal) are calculated as disagreements. Hence, all disagreements count equally strong in the calculation (although there exist methods that take into account different types of disagreements (cf. (Krippendorff, 1980)), such a method does not exist for categories from ontological hierarchies yet). If we factor out and align these occurrences in the results, agreements improve up to 90%. An interesting finding from the annotation evaluation is also that some annotation examples elucidate dependencies from the modality hierarchy, without explicit knowledge of the annotators as they were not explicitly informed about the hierarchical structure of the modalities in GUM-Space. In the sentence "Es ist gegenüber von mir. [It is opposite of me.]" from the German sample, for instance, the annotation of the 'golden standard' is Proximal as the modality for the relation *gegenüber* (*opposite*). Both German annotators, however, annotated this modality as a FrontProjectionExternal. Although this category is slightly too specific for the relationship of being on the opposite side, as the locatum does not necessarily have to face the relatum, it is a subcategory of Proximal in the modality hierarchy of GUM-Space. Hence, both categories show a strong connection, which is formalized by the hierarchical relationship in GUM-Space and implicitly indicated by the annotations.

Another result from the inter-annotator agreement study is that alternative readings are often ignored. In particular, examples such as "Walk down the street." or "Drive down that road." can in principle either indicate that the path of the motion follows the reference object (the "street" or the "road"), which is annotated as PathRepresentingInternal in GUM-Space, or it can indicate that the path of the motion also decreases in a vertical direction, annotated as SpecificDirectional. While one English annotator used the second annotation category consistently throughout the sample data phase, only the second English annotator sometimes annotated both alternative readings in the spreadsheet document. Again, given our strict way of measuring the agreement, the missing annotation of the alternative reading of the first annotator caused disagreements in the calculation.

In general, however, no systematic differences can be found with regard to disagreements between annotators. Besides the differences described here, most disagreements are caused by typing errors (e.g., annotators use LeftProjection although the sentence contains a RightProjection "to the right of") or accidental and random errors.

5. Discussion

In this paper, we argue that inter-annotator agreement can be used to evaluate the adequacy of linguistic ontologies,

[4]The annotations were calculated by using the ReCal tool, available at `http://dfreelon.org/utils/recalfront`

in particular GUM-Space. The presented inter-annotator agreement study shows encouraging results that GUM-Space is able to structure spatial language in an adequate way. Another promising result is that annotators are able to understand and use the complex annotation schema of GUM-Space. For future research, we will conduct further annotation studies containing more sample data, in particular, balanced samples that contain all the different categories. Also, calculating agreement for similar but slightly different annotations, i.e., those categories that are specified hierarchically close together, have to be considered in an appropriate way. A fine-grained distinction between major and minor disagreements, for instance, based on similarity, might be more appropriate for measuring inter-annotator agreements.

Moreover, annotators are not going to be asked to clean up the sample, i.e., removing the non-spatial data as requested from the current manual, because it can lead to different interpretations in the annotation of different annotators. These results can then not be used for the calculations. In our case study, we had to ignore 15 English and 16 German configurations because of different results from the clean up task.

In general, 100 sample sentences are too few examples to calculate a thorough inter-annotator agreement, although they resulted in 115 English and 136 German configurations. However, we were able to show that (1) annotators are able to learn the complex GUM-Space instructions by using a manual and (2) that our first inter-annotator study shows promising results. For our next inter-annotator study, we currently collect 300 new sentences to build a balanced data sample. It contains at least five example sentences for each modality category and a huge variation for motion, orientation, and locating configurations together with possible modifications. Training and supervised data samples will be taken from the material presented in this paper.

Acknowledgements

We gratefully acknowledge the financial support of the Deutsche Forschungsgemeinschaft through the Collaborative Research Centre SFB/TR 8 Spatial Cognition. The author would like to thank the four annotators M. Mathebula, D. Nakath, A. Scholz, and S. Wilkinson, and E. Andonova and J. Bateman for fruitful discussions.

6. References

John A. Bateman, Renate Henschel, and Fabio Rinaldi. 1995. Generalized Upper Model 2.0: documentation. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

John A. Bateman, Thora Tenbrink, and Scott Farrar. 2007. The role of conceptual and linguistic ontologies in discourse. *Discourse Processes*, 44(3):175–213.

John A. Bateman, Joana Hois, Robert Ross, and Thora Tenbrink. forthcoming. A linguistic ontology of space for natural language processing.

Janez Brank, Marko Grobelnik, and Dunja Mladenić. 2005. A survey of ontology evaluation techniques. In *Proceedings of the 8th International Multi-Conference on Information Society (SIKDD'2005)*, pages 166–169.

Balakrishnan Chandrasekaran, John R. Josephson, and V. Richard Benjamins. 1999. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, pages 20–26.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Kerstin Fischer. 2007. The role of users' concepts of the robot in human-robot spatial instruction. In Thomas Barkowsky, Markus Knauff, Gerard Ligozat, and Dan Montello, editors, *Spatial Cognition V: Reasoning, Action, Interaction*, pages 76–89. Springer, Berlin, Heidelberg.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. Stataxis Publishing Company.

Michael A. K. Halliday and Christian M. I. M. Matthiessen. 1999. *Construing experience through meaning: a language-based approach to cognition*. Cassell, London.

Peter A. Heeman and James Allen. 1995. The trains 93 dialogues. Trains Technical Note 94-2, Computer Science Dept., University of Rochester, mar.

Joana Hois and Oliver Kutz. 2008a. Counterparts in Language and Space – Similarity and $\mathcal{S}$-Connection. In Carola Eschenbach and Michael Grüninger, editors, *Formal Ontology in Information Systems (FOIS 2008)*, pages 266–279. IOS Press.

Joana Hois and Oliver Kutz. 2008b. Natural language meets spatial calculi. In Christian Freksa, Nora S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Proceedings of International Conference Spatial Cognition VI*, pages 266–282. Springer.

Boris Konev, Carsten Lutz, Dirk Walther, and Frank Wolter, 2009. *Formal Properties of Modularisation*, pages 25–66. Springer-Verlag, Berlin/Heidelberg.

Klaus Krippendorff. 1980. *Content Analysis - An Introduction to Its Methodology*. Sage Publications, California.

Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and Ewan Klein. 2001. Training Personal Robots via Natural-Language Instructions. *IEEE Intelligent Systems*, pages 38–45.

Stephen C. Levinson. 2003. *Space in language and cognition: explorations in cognitive diversity*. Cambridge University Press, Cambridge.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604.

Leo Obrst, Benjamin Ashpole, Werner Ceusters, Inderjeet Mani, Steve Ray, and Barry Smith, 2007. *The Evaluation of Ontologies - Toward Improved Semantic Interoperability*, pages 139–158. Springer-Verlag, New York.

Robert Ross, Hui Shi, Tilman Vierhuff, Bernd Krieg-Brückner, and John Bateman. 2005. Towards dialogue based shared control of navigating robots. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bern-

hard Nebel, and Thomas Barkowsky, editors, *Proceedings of International Conference Spatial Cognition IV*, pages 478–499, Berlin, Heidelberg. Springer.

Robert J. Ross. 2008. Tiered models of spatial language interpretation. In Christian Freksa, Nora S. Newcombe, Peter Gärdenfors, and Stefan Wölfl, editors, *Spatial Cognition VI. Learning, Reasoning, and Talking about Space*, pages 233–249. Springer.

Hui Shi and Thora Tenbrink. 2009. Telling rolland where to go: HRI dialogues on route navigation. In Kenny Coventry, Thora Tenbrink, and John Bateman, editors, *Spatial Language and Dialogue*, pages 177–189. Oxford University Press, Oxford.

Leonard Talmy, 2006. *The fundamental system of spatial schemas in language*, pages 37–47. Mouton de Gruyter, Berlin.