On the Representation of Perceptual Knowledge for Understanding Reference Expressions

Philipp Spanger and Takenobu Tokunaga

Tokyo Institute of Technology
Graduate School of Information Science and Engineering
Department of Computer Science
{philipp,take}@cl.cs.titech.ac.jp
http://tanaka-www.cs.titech.ac.jp/jp/index.html

Abstract. Recent research has enabled important progress in developing agents aimed at real-world linguistic interaction with humans. Hence, within the general shift of research focus from "information" to "knowledge", an important question is how to apply large-scale knowledge resources in order to improve agents' capabilities of linguistic interaction with humans. This paper presents research toward an efficient representation of the necessary perceptual knowledge in dialogue with a particular focus on reference expressions. We generalize an existing formal model of reference expressions involving perceptual grouping in order to account for a number of types of reference expressions that the previous model could not account for. Our model yields an increase in both coverage and accuracy of referent identification − which has been confirmed in preliminary experiments. We outline an algorithm for the future application of this model to other languages, showing how the model can be extended to deal with large-scale multi-language input data.

Keywords: representation of perceptual knowledge, perceptual grouping, reference expressions, language-independent systems.

1 Introduction

Recently, the utilization of large-scale knowledge resources (LKR) has been a central issue in achieving progress in different research areas such as analysis of spoken language characteristics, systematization of archeological information or language-learning support systems. In particular, the application of LKR in research in the field of linguistics is a very promising research direction. At the same time, developments in a multitude of research areas like speech recognition, robotics, etc. have enabled important progress in developing agents aimed at real-world interaction with humans. Thus, within this general shift of research focus from "information" to "knowledge", an important question is how to use large-scale knowledge resources in order to improve agents' capabilities of interaction with humans through natural language. An important research aim in improving agents' capabilities of interaction with humans has been to improve

T. Tokunaga and A. Ortega (Eds.): LKR 2008, LNAI 4938, pp. 280–294, 2008.

their natural language understanding. A fundamental type of human expression − in particular in task-oriented dialogue − are reference expressions. This type of expression is a linguistic entity used to discriminate a specific object from its environment and the rest of the world.

Thus, an agent's capability to handle this type of linguistic expression correctly is an important part of increasing human-agent interaction capabilities. Reference expressions are to a large degree multi-modal; i.e. they include exophoric expressions such as "this one" or "that" in connection with gesturing (e.g.; pointing). It is clear a fuller model of reference expressions must be a multi-modal model including an account of these different channels and how they combine ([10]). As a first preliminary step towards this aim, we intend to generalize a current model of reference expressions limited to the linguistic channel as a basis for future application in a multi-modal environment. Hence, in this paper the term "reference expression" refers to a single-channel linguistic expression which moreover includes no anaphora and is functioning as a full description for identifying objects in the world such as "the blue ball in front of the table". There has been significant research in the area of how the human cognitive process − and thus also knowledge of the world − and language production and understanding are linked. Specifically, cognitive linguistics researches in this area in a wide variety of directions ([4], [5]). Generally, it is clear human language understanding/production is directly linked to a human's general world knowledge as well as the knowledge acquired/exchanged in a particular dialogue.

A still unsolved problem is how world knowledge, including linguistic knowledge is represented in the human brain. In order to enable agents to effectively communicate with humans through natural language, a critical task is to provide an effective model of the knowledge as applied by humans in the course of diaolgue. Fundamentally, the knowledge used by a human to comprehend a certain linguistic expression can be separated into two types of necessary knowledge; 1) general linguistic knowledge (i.e. as encoded in a grammar and vocabulary) and 2) world knowledge. In the particular case of understanding reference expressions, the necessary knowledge about the world can be separated further into a) general world knowledge as well as b) particular perceptual knowledge for this particular situation. In this case, we broadly define "perceptual knowledge" as comprising all knowledge generated through human perception of a specific situation; e.g. the location of the objects in the domain, their colour, shape, etc and their respective relations to each other. This knowledge is utilized by humans to produce and understand linguistic reference expressions. Obviously, the human perceptual apparatus is capable of extracting a potentially intractably large amount of perceptual knowledge from any given specific real-world situation and a key problem is to decide which of this information is relevant in which domain for the given task of language understanding. The majority of previous work on linguistic reference to a target-object among other distractors, (e.g. [1], [2], [3], [11], [9]) utilized perceptual knowledge of the attributes of the target and binary relations between the target and distractors, using surface differences of the objects. The Incremental Algorithm [3] is an important example of this kind

of algorithm. These works mostly deal with developing algorithms for the generation of natural language reference expressions that work sufficiently well in domains where the objects and distractors have a significant surface difference.

However, there is a significant case of failure within this general framework. In case no significant surface difference and no binary relation useful to distinguish the target from the distractors exists, such methods cannot generate a natural linguistic expression enabling hearers to identify the target. Furthermore, these methods cannot provide a model to understand any linguistic expression generated by humans in such a case. This paper seeks to contribute to research in the area of understanding of reference expressions in such a domain. Previous research has underlined the importance of perceptual grouping in understanding [12], and generating [6] reference expressions. Perceptual grouping is defined as the human ability to recognize similar objects, or objects in close proximity to each other. Effective understanding of human reference expressions in this specific domain requires recognition of similar or proximal objects, i.e., perceptual grouping, and requires making use of n-ary relations among objects in each recognized group. Research based on this understanding has produced comparably good results in both the understanding and generation of reference expressions. While this general approach has proven valuable in both the understanding and generation of this type of expression, it has been hampered − both theoretically and in practice − by a strong limitation on the type and structure of expression. That is, it has been assumed reference expressions exclusively apply a linear process of narrowing down of the referent (represented by a "Sequence of groups-representation" (SOG) in [12] and in its generalized form in [6]). However, this means other relations between sets of perceptual groups (appearing in reference expressions) like intersection or subtraction cannot be represented.

Data of experiments in several languages (Japanese, English, German, French) indicate that while the overwhelming majority of expressions (in all four languages) is based on this type of process, it is far from the only or even always the most natural one for humans. In particular, humans are capable and in some cases prefer to refer to different types of relations between sets of similar objects using either intersection or subtraction. In certain cases this simplifies the expression significantly or is more natural. For example, in Figure 1, through use of the expression of "ignoring the three balls in the right back", the subject applies a subtraction-relation between the group of all balls in the domain and the "three balls in the right back". The target object is referred to from within the remaining set. This is one example of a process of referring, that cannot be represented in the previous model. Hence, in order to develop the promising framework of application of a representation of perceptual grouping in reference expressions, it is necessary to generalize the existing model such that it can accomodate these more complex cases. This paper tackles this task. This will make a contribution to increasing our understanding of the necessary representation of perceptual knowledge for the efficient understanding of reference expressions in human-agent linguistic interaction. Furthermore, it will provide a general theoretical model of this type of expression.

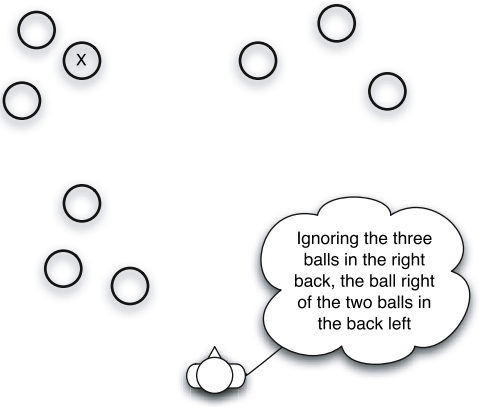

Fig. 1. An example of a reference expression using subtraction

For the overall task of constructing an efficient LKR for application in the domain of human-agent linguistic interaction, the question of how to represent the necessary knowledge, including perceptual knowledge, is critical. A solution to this task will also enable in the future an efficient systematization of this LKR. Through research into the question of how to represent perceptual knowledge for the understanding of reference expressions, our work seeks to contribute towards this aim. As in [12], we consider here that understanding reference expressions consists of two stages: (a) semantic analysis, i.e., analyzing expressions to extract semantic information, and (b) referent identification, i.e., discriminating referents by using extracted information. Below, we describe the proposed generalized model for perceptual knowledge for understanding of reference expressions. We will explain how this model handles the more complex cases the previous model could not deal with. We also explain some modifications of the algorithm for construction of a representation of perceptual grouping proposed in [12], as necessitated by the more general model proposed in this paper. We will discuss the collected data (in English as well as other languages) and the implementation of the proposed model in a simple prototype that yielded an increase in both coverage and referent identification. Finally, we will give an overview of future work on this topic.

2 COG (Combination of Groups): A Formal Model of Reference Expressions

As stated above, previous work on reference expressions focused on using surface differences of the objects. However, in the case of the absence of significant surface difference and if no binary relation useful to distinguish the target from the distractors exists, such methods failed.

To solve this insufficiency, [12] proposed a method of generating Japanese reference expressions that utilizes n-ary relations among members of a group. Necessarily this framework included a representation of the perceptual grouping

process − called an intermediate "Sequence of groups" (SOG) − representation. This representation captured the linear process of narrowing down of the referent. However, their framework only dealt with the limited situations where exclusively homogeneous objects are randomly arranged (as in Figure 1). Thus, the representation in their method could only be applied in the case of spatial n-ary relations, and could not handle attributes and binary relations between objects which have been the main concern of past research.

2.1 The (Extended) SOG Representation and Its Limitations

We outline here the extended SOG-model and point out its limitations.[12] assumed a situation with randomnly-arranged homogenous objects and focused exclusively on the representation of the spatial subsumption relations between consecutive groups. Thus, the intermediate representation of perceptual knowledge they proposed between a reference expression and the situation that is referred to by the expression, did not explicitly denote relations between groups in the original SOGs (as shown below).

$SOG : [G_0, G_1, ...G_n]$
$G_i : a\ group$

In order to take into account other types of relations between groups, [5] proposed then an extended SOG representation where types of relations are explicitly denoted as shown below.

$SOG : [G_0 R_0 G_1 R_1 ... G_n]$
$G_i : a\ group$
$R_i : relation\ between\ G_i\ and\ G_{i+1}$

This extended representation accounts for two types of relations between perceptual groups: intra-group relations and inter-group relations. Of course, for any intra-group relation, by definition, G_i subsumes G_{i+1}, that is, $G_i \supset G_{i+1}$. Intra-group relations are further classified into subcategories according to the feature used to narrow down G_i to G_{i+1}. In this model, in case R_i is an inter-group relation, G_i and G_{i+1} are mutually exclusive, that is, $G_i \cap G_{i+1} = \emptyset$. However, this leaves out cases of other inter-group relations, in particular other combinations of perceptual groups like intersection ($G_i \cap G_{i+1}$) or subtraction ($G_i \setminus G_{i+1}$). The necessity of incorporating this type of expression can be demonstrated by, for example, Figure 1 (example from the collected expressions).

2.2 The COG (Combination of Groups) Representation

We propose the Combination of groups (COG) - model as an efficient representation of the perceptual grouping process and demonstrate how it resolves the limitations of the previous approaches. We provide an example analysis in the COG - model of an expressions that the previous model could not handle. The COG - representation is a generalization and extension of the (extended) SOG-representation. Its flexible order of grouping ("linear" and "non-linear")

better captures the natural variety of human reference expressions. It includes the SOG-representation as a special case. The initial SOG-model was extended by [6] to the case of different relations among sets, which arise in more complex environments with a number of different objects. However, both the extension as well as the initial model share the same weakness in not accounting for reference expressions that include in some form a non-linear process of narrowing down the referent. In this context, we denote by non-linear process any process that involves reference to groups of objects, where these different groups are related to each other neither by a simple subsumption-relation nor by a simple inter-group relation where the groups share no elements. This means the relation between the sets is not a subsumption-relation. Furthermore, the intersection of the two sets is neither empty nor equals any of the two sets.

Thus, in order to improve the model we implemented the generalization that is able to handle these cases. In order to demonstrate the basic validity of the proposed generalization, we based our implementation on the simpler earlier model. However, as the initial as well as the extended SOG-model shared this same weakness, we note that this generalization is applicable as well for the extended SOG-model. [12] point out that in their method, "most errors in semantic analysis are due to non-linearity of referring". This is because the SOG-model presupposes linearity in referring and thus cannot handle these cases. We recall that linearity in this case refers to the fact that between subsequent perceptual groups exclusively subsumption-relations exist. We conducted a preliminary data collection experiment. The analysis of the collected data indicated two general cases of "non-linearity" in referring, i.e. of the existence of relations different from the subsumption-relation. We noted subjects tended to use either the intersection or the subtraction − relation. Hence we concentrated on the implementation of these two relations. Other more complex combinations of these relations are possible, but they can be reduced to a combination of these two more simple set-relations. As the name "Sequence of groups" indicates, this previous model has a "flat" structure, i.e. it only accounts for one type of binary relation to the immediately preceding group: the subsumption relation. In fact, the subsumption relation is simply a special case of the intersection relation. In contrast, the proposed COG-model allows an internal structure within the representation of the perceptual process; i.e. a reference to any previous group or combination of groups. Thus, this model can correctly represent the cases where the SOG-model fails (i.e.: subtraction and intersection relation). The general COG-model includes the following relations where G_n denotes a group or combination of groups.

(1) intersection of groups : G_i and G_j with the result G_k: $G_i \cap G_j = G_k \neq \emptyset$

$$\left.\begin{matrix} G_i \\ G_j \end{matrix}\right] \xrightarrow{\cap} G_k$$

(2) subtraction of groups : G_i and G_j with the result $G_k : G_i \setminus G_j = G_k \neq \emptyset$

$$\left.\begin{matrix} G_i \\ G_j \end{matrix}\right] \overset{\setminus}{\longrightarrow} G_k$$

(3) subsumption relation : $G_j \subset G_i$

$$G_i \longrightarrow G_j$$

(4) inter-group relation : $G_i \cap G_j = \emptyset$

$$G_i \Rightarrow G_j$$

We can see that while the SOG-model (including in its extended form) only allows unary relations, in our model we include representation of binary relations between perceptual groups: the intersection as well as the subtraction relation. Thus, theoretically an arbitrary level of complexity of grouping-(reference) expressions can be represented in this model. It is quite easy to extend our model to relations like:

$G_i \overset{colour}{\longrightarrow} G_j$,
$G_i \overset{size}{\longrightarrow} G_j$,
$G_i \overset{type}{\longrightarrow} G_j$.

If we incorporated these types of relations into the COG-model, our model would also be able to deal with the relations handled by the extended SOG-model and thus include the extended SOG-model as a special case. In the following, we analyze a characteristic example that cannot be accounted for by the previous model (displayed in Figure 2)). We provide the analysis of this example based on our proposed model in detail and then show how and why its analysis fails in the previous model.

2.3 Example Analysis in COG

In this section we present an analysis of an expression the previous model cannot handle, but that the COG-model can appropriately analyse. This will demonstrate how representation of perceptual knowledge in the COG-model is useful for application in the domain of reference expression understanding. Furthermore, we will explain how the previous model can be reduced to a special case of the proposed COG-model. We provide the analysis of a reference expression employing a subtraction- relation (see Figure 2) between different perceptual groups. As we pointed out previously, this cannot be handled by the previous model.

We recapitulate the phrase: “Ignoring the three balls in the left back, the ball in the back”. The analysis in the COG-model would look as follows.

$$\begin{matrix} \{a,b,c,d,e\} \\ \{a,b,c,d,e\} \longrightarrow \{c,d,e\} \end{matrix}\Bigg] \overset{\setminus}{\longrightarrow} \{a,b\} \longrightarrow \{b\}$$

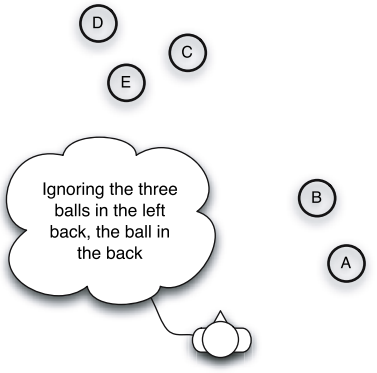

Fig. 2. An example of a reference expression including subtraction-relation

First, from the set of all balls $\{a, b, c, d, e\}$, the user selects "the three balls in the left back" $\{c, d, e\}$. Thus using the subtraction-relation, from the set of all balls $\{a, b, c, d, e\}$ in the domain, the user focuses the attention on the remaining set of objects $\{a, b\}$, from which the target $\{b\}$ is then selected by using a simple subsumption-relation. Here it is important to note that the cardinality of the group of objects of the result of the subtraction-relation $\{a, b\}$ is not explicitly referred to. As we noted previously, in the SOG-model, there is no way to represent this relationship between $\{a, b\}$ and $\{a, b, c, d, e\}$ Thus, if we were to try to provide an analysis based on the SOG-model, this would lead to a representation as follows.

$SOG : [\{a, b, c, d, e\}R_0(\{c, d, e\}R_1\{a, b\})R_2\{b\}]$
$G_0 = \{a, b, c, d, e\}; G_1 = \{c, d, e\}$
$G_2 = \{a, b\}; G_3 = \{b\}$
In particular we note that the following holds: $G_1 \cup G_2 = G_0$.

This type of relation between immediately succeeding groups ($G_1 \cup G_2 = G_0$) cannot be represented in the SOG-model. Hence, we see that the SOG-model fails in this case. We note the SOG-model by [12] is a special case of the proposed COG-model, in the case that all relations are restricted only to subsumption-relations and thus only relations to immediately preceding groups are allowed. This means if for all groups G_i it holds that $G_i \supset G_{i+1} \supset G_{i+2}...$, the COG-model goes over into the simple SOG-model.

2.4 Perceptual Grouping

As pointed out previously, the algorithm for perceptual grouping proposed in [12] only recognizes groups that are explicitly referred to, with their cardinality specified (e.g.: "the three balls in the ..."). As pointed out above, in addition to this type of perceptual grouping, humans carry out perceptual grouping by exclusion, i.e. "the balls right to the table and ...". Here the subject forms a specific perceptual group G_i , without specifying a cardinality of the group. Thus,

we implemented a simple but conceptually important modification to [12] that handles these cases. This modification was important particularly in order to implement the more complex combinations of groups like repeated intersections and complements. In fact, the limitation of allowing only perceptual groups with explicit cardinality is closely connected to the linear structure of the SOG-model; since neither intersection nor subtraction-relations on sets were permitted, the only way to select a set of certain elements from a super-set is to specify its cardinality. Hence, in our generalization to allow a wider range of set-relations, we needed to implement a concomittant generalization of the perceptual grouping algorithm proposed in [12].

3 Implementation

Following [12], we consider the general process of reference expression understanding to consist of the following two stages (a) semantic analysis, i.e., analysing expressions in order to extract semantic information, and (b) referent identification, i.e. uniquely recognizing referents by using the extracted semantic information. Generally, the methods of [12] are employed for both stages, in particular in the process of perceptual grouping. However, in both stages some modification of the methods were implemented.

3.1 Semantic Analysis

A critical task is the extraction of the relevant information from the linguistic expressions. [12] used a simple pattern matching-technique for extracting the necessary information from the linguistic expressions instead of full parsing. In this paper, we used the Stanford Parser (see [8]) to extract a basic syntactic structure based on PCFG (probabilistic context-free grammar) ([7]). This improvement lays the basis for building an LKR of syntactic structures and associated with it the necessary information to be extracted for its use in understanding reference expressions. Information mining and machine learning techniques could then be applied to facilitate the use of this LKR in the area of human-agent linguistic interaction. The utilized framework of PCFG is a context-free grammar in which every production is assigned a probability. The final probability of any parsing of a specific sentence is calculated as the product of the probabilities of all the productions in a specific parse. The parse with the highest probability is then selected by the stochastic grammar.

We then analysed the basic syntactic structures of the user inputs and recognized that to a large degree the syntactic structure gives a good clue as to how to separate a clause into groups for extracting the required information. In order to elucidate this process, we give an example in the following:

User input sentence: “the rightmost ball among the three balls at the back left”. The raw parser output is as follows:

(ROOT (NP (NP (DT the) (JJ rightmost) (NN ball)) (PP (IN among) (NP (NP (DT the) (CD three) (NNS balls)) (PP (IN at) (NP (DT the) (JJ back) (NN left)))))))

Represented as syntactic tree it looks as shown in Figure 3.

The representation in COG would look as follows:

$\{$a,b,c,d,e,f,x$\}\longrightarrow\{$a,b,x$\}\longrightarrow\{x\}$.

Our basic aim is to get from the surface linguistic form to the forming of perceptual groups corresponding to the formal COG-representation. As the initial group $\{a, b, c, d, e, x\}$ is not expressed explicitly, we basically look for how the following part of the above COG-representation is reflected in the syntactic structure.

$$\{a, b, x\} \longrightarrow \{x\}.$$

In the example above, we see that the perceptual group $G_2 = \{x\}$ is represented by the NP_A which corresponds to the linguistic surface expression "the rightmost ball". The PP_B corresponds to the $G_1 = \{a, b, x\}$. We need to extract the relation R_1 between these two groups $G_1 R_1 G_2$. This relation can be extracted from the first component of PP_B which is IN_C. The label IN_C corresponds to "among" in the linguistic surface form, which indicated a subsumption-relation. The NP_D includes the cardinality of the group "three" and the location: "back left". Japanese is a head-final language and hence the

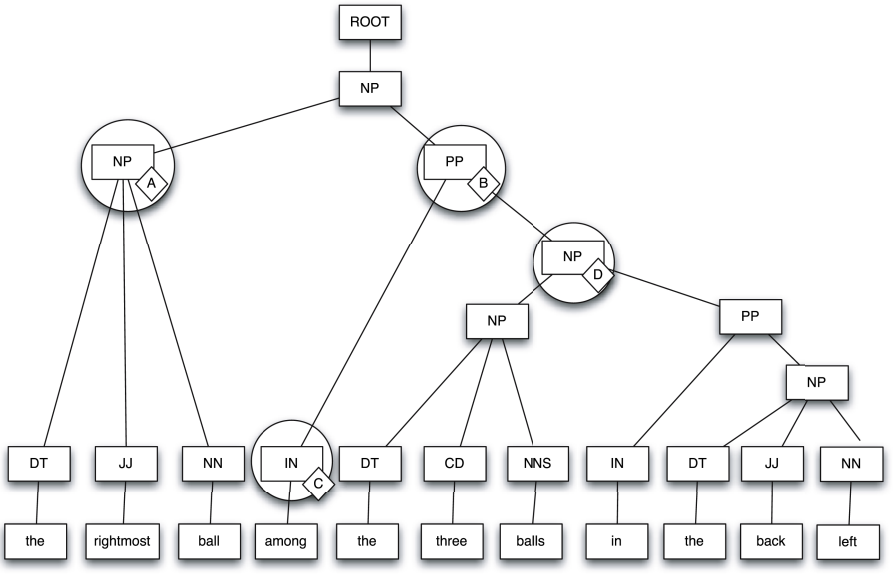

Fig. 3. An example of syntactic structure of user input

order of sub-expressions in the linguistic surface-form has the same order as the groups in the SOG and COG representation. This however is obviously different in languages such as English which are not head-final. In our example case here, the order of sub - expressions is exactly opposite to the order of groups in the relevant part of the COG-representation. We note that in both Japanese (as reported by [12]) as well as in English, the total set is generally not mentioned. From the syntactic structures collected, there were several regularities which seem to indicate a fundamental connection between syntactic structure and the process of perceptual grouping. However, we utilized the observed regularities in a simple ad-hoc manner for semantic analysis, as the main aim of this work was to evaluate the proposed COG-model. We acknowledge this is still a very partial progress to the pattern matching technique employed in [12]. In future work, a more complete analysis and comparison of different grammatical approaches should be carried out. In particular, a study of a large amount of testdata could be the basis for applying learning algorithms over the syntactic structures in order to enable a more systematic account of the connection between syntactic structures and the process of perceptual grouping. Based on this work, a deeper understanding of how the perceptual process of forming groups is reflected in the syntactic structure can be gained. This in turn could then be applied to improving our representation of perceptual knowledge necessary for human-agent linguistic interaction.

3.2 Referent Identification

As the next step following the semantic analysis process, we carried out referent identification by using the extracted information. (It was assumed all the participants in a specific situation shared an appropriate reference frame.) The process of the identification of the referent goes from left to right. The members of each group are identified as a result of the application of a referent identification algorithm. The algorithm applied to G_i depends on the relation between G_i and G_{i+1}. Each algorithm is fundamentally an identification function with a set of objects and information to specify referents as arguments. As pointed out previously, we fundamentally used the algorithm for perceptual grouping as proposed in [12].

4 An Outline of a Language-Independent Algorithm

An important issue in the future for the construction and design of LKR in the linguistic domain is its application to a multi-lingual domain. Furthermore, the question we are concerned with here − namely the effective representation of perceptual knowledge − should be studied in a multi-lingual environment in order to discover commonalities as well as differences over the different languages. In order to facilitate this future work, we outline here how our model can be implemented in other languages. The proposal of the COG-representation is based on observations of a data - collection in English, French and German. It

captures the general structure of reference expressions in these languages. Of course, in order to implement a system of understanding reference expressions comparable to the system we prepared for English, the possibly very significant differences in syntax have to be accounted for. From the test-data, our observation is that there is no significant difference in the process of perceptual grouping that would force a fundamental revision of the algorithm proposed in this paper. However, we found some tendencies of preference of certain types of expressions, which differed over the different languages. Generally, we observed some interesting characteristics in the collected data in the different languages. German expressions showed a significant variation in the syntactic structure of the expressions (as well as the used vocabulary), while the expressions supplied by the French subjects showed a very high degree of similarity of syntactic structure. Further study and examination of data in other languages should illuminate this phenomenon and the connection between cognitive and linguistics processes. In fact, clarifying the interdependence between cognitive and linguistic processes and how they differ over different languages would provide an understanding that could be critical in many areas of natural language processing. In order to implement and test the proposed model in other languages, in particular the following modules should be prepared:

- Syntactic parser
  The output of the parser should be analysed for indicators in the syntactic structure of distinct grouping. The analysis of the syntactic structure of the input expression forms the basis for further information extraction.
- Information extraction-module
  Based on the previous step, words/ syntactic structures indicating a particular set-relation of perceptual groups (e.g.: subtraction - relation in English "ignoring :." etc.) should be identified and applied to extract the relevant information.

In our implementation of the methods of referent identification developed in [12] for Japanese, we noted there was no significant modification necessary for the application of these methods to English, other than those explained above. Thus, our data-collection experiment indicates a very universal process of perceptual grouping. The testdata in English and Japanese − being two languages with significant differences in syntax − provide at least a good basis for making this hypopthesis. Our testdata in French and German confirm this hypothesis. Of course the amount of testdata is very small and thus these hypotheses need to be tested using a more comprehensive set of data.

5 Evaluation

We implemented the proposed model in Java and applied it to the expressions collected in our data-collection experiment. We then evaluated the referent identification accuracy of the proposed model.

5.1 Experiment

We carried out a data-collection experiment for English, where we provided the 12 different arrangements of balls in a 2-D bird's-eye image to the subjects (taken from the appendix in [12]). 12 subjects whose native language is English participated in the experiment over the internet. They were provided an arrangement with the choice to either input an expression they felt appropriate or to abandon this specific arrangement in case subjects were not able to think of an appropriate reference expression. This should have produced 144 expressions, however 7 judgements were abandoned and 15 expressions were either nonsensical or obviously insufficient to identify the referent. Hence, we obtained 122 English reference expressions. [12] referred to about 8% of Japanese collected expressions that included non-linearity of referring. In the English data collected in our experiment, we noted a 7% frequency of reference expressions that include non-linearity. This points toward very similar frequency of this type of expression. However, the amount of the English data in particular is small; hence in order to confirm this hypothesis a larger data set is necessary.

5.2 Results

We did not have the previous system (in Japanese) at our disposal. Thus we implemented the algorithm as outlined in [12] in English. This was in order to provide a baseline for our proposed enhanced model. The result of this system is represented in Table 1 in comparison with the results of the Japanese system, displayed in Table 2. Our system in English based on the algorithm in [12] gives largely comparable results to the system of [12] for Japanese. The slight decrease in accuracy (about 2%) can be attributed in part to a lack of fine-tuning of the algorithms for perceptual grouping.

Table 1. Results of previous model in English

Expression	*Pop.*	*Ident.*
Total	122	77.0%
Applicable	107	83.1%

Table 2. Results of previous model in Japanese

Expression	*Pop.*	*Ident.*
Total	476	78.8%
Applicable	425	84.7%

Table 3. Results of COG - model in English

Expression	*Pop.*	*Ident.*
Total	122	82.6%
Applicable	114	89.2%

The result of the implemented system based on [12] in English is represented in Table 1. It shows that the simple implementation in English yielded a comparable result to the Japanese system, while having slightly less accuracy. This might be attributed to slight differences in implementation; e.g. setting of some parameters in the formulas of the perceptual grouping methods. We then implemented the GOG-model and the result is represented in Table 3. This implementation of the GOG-model yielded an increase of 5.6% in comparison to the SOG-model in English. The final accuracy achieved for English was 82.6%.

5.3 Error Analysis

From the collected expressions, there were two significant types of errors that are described in the following.

(a) Errors in semantic analysis
There were two main types of errors in semantic analysis. One type of expression referred to a particular part of the body of the person in the picture as referent, e.g.: "the one in front of my left shoulder". The other frequent type of expression that cannot be handled by our system are expressions that refer to an action, like : "Take away the two left ones and you'll have it now as the most left ones". There were 3 expressions of this type. This type of expression appeared in all three languages of our experiment, thus indicating it is not an isolated phenomenon. A future system should be able to handle expressions of this type involving actions.

(b) Referent identification
In the main, errors were due to reference to geometric forms − in particular lines − and our current system cannot handle any perceptual grouping involving this type of figure. We acknowledge that this is a preliminary evaluation, as the test-data is less than a quarter of the amount of expressions collected in the other system.

6 Conclusion

In the framework of research into efficient representation of perceptual knowledge in language understanding, we proposed a generalized model of reference expressions that seeks to capture the varied forms of reference expressions employed by humans. We proposed a generalization of a previous model, of which the previous model is a special case. We demonstrated how our proposed model can handle several types of expressions that the previous cannot handle. We then implemented our model. We measured both an increased coverage and increased identification accuracy in comparison to the previous model in English by 5.6% to a total identification accuracy of 82.6%. We reported some observations on the collected data in other languages in comparison and gave an outline of how to implement this general framework in other languages. Our model has so far only been implemented in the area of understanding of reference expressions, but

it should be noted that it could be extended to the generation of reference expressions. Furthermore, the construction of an LKR, comprising languages other than English and in particular also a syntactic representation of the linguistic input is a critical task in the future. The proposed model is simply a linguistic model of reference expressions in a 2-D environment. In order to increase the efficiency of human-agent communication it is necessary to incorporate other channels of communication ("multi- modality") and combine the information of these. In the future, we plan to extend and adapt the proposed model in this thesis to a multi-modal environment. The construction of multi-modal LKR for application in the human-agent interaction domain in the future is an important goal and we see our work as a step towards realizing this.

References

1. Appelt, D.: Planning english referring expressions. Artificial Intelligence 15(3), 143–178 (1985)
2. Dale, R., Haddock, N.: Generating referring expressions involving relations. In: Proceedings of EACL, Berlin, pp. 161–166 (1991)
3. Dale, R., Reiter, E.: Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science (PDF) 19, 233–263 (1995)
4. Kristiansen, et al (eds.): Cognitive linguistics: Current applications and future perspectives. Mouton de Gruyter, Berlin, New York (2006)
5. Vyvyan, Evans, B.B., Zinken, J.: The cognitive linguistics reader. Equinox, London (2007)
6. Funakoshi, K., Watanabe, S., Tokunaga, T.: Group based generation of referring expressions. In: Proceedings of the Fourth International Natural Language Generation Conference, pp. 73–80 (2006)
7. Johnson, M.: Pcfg models of linguistic tree representations. Computational Linguistics 24(4), 613–632 (1998)
8. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003)
9. Krahmer, E., van Erk, S., Verleg, A.: Graph-based generation of referring expressions. Computational Linguistics 29(1), 53–72 (2003)
10. Kranstedt, A., et al.: Deictic object reference in task-oriented dialogue. In: Rickheit, G., Wachsmuth, I. (eds.) Situated Communication, pp. 155–207. Mouton de Gruyter, Berlin (2006)
11. van Deemter: Generating referring expressions: Boolean extensions of the incremental algorithm. Computational Linguistics pp. 28 (1), 37–52 (2002)
12. Watanabe, S., et al.: Understanding referring expressions involving perceptual grouping. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)