Learn to Speak Like Normal People Do: The Case of Object Descriptions

Michael Zock[1], Guy Lapalme[2] and Mehdi Yousfi-Monod[3]

[1]Laboratoire d'Informatique Fondamentale (LIF), CNRS, UMR 6166
Case 901 - 163 Avenue de Luminy, F-13288, Marseille Cedex 9, France
[2]RALI-DIRO, Université de Montréal
C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada
[3]Research and Development, NEIIO, Montpellier, France

Abstract. Successful communication requires loads of various knowledge including peoples' habits to express their thoughts. We deal here with the description of objects composing a scene. This is called 'reference generation', a task very frequently performed in language production. The skill of language production requires at least two kinds of competencies: one dealing with the mapping of concepts to forms (linguistic competency), the other dealing with the choice of the various resources (communicative competency). Messages can be expressed in many ways and at various levels: lexical, grammatical, morphological, etc. Since different means tend to produce different effects, students have to learn when to use which. This kind of knowledge, commonly called 'pragmatic knowledge' or 'communicative competency', can rarely be formalized as solid rules. It is largely based on experience. People learn on the basis of correlations, that is they realize that changes of the situation may reflect in language: different inputs (ideas, objects of a scene) yielding different outputs, i.e. linguistic forms. We present here a setting that allows for this kind of learning. It is a web-based application that generates a scene and various descriptions of its components. Users can change the scene and watch how these choices affect (or not) the linguistic form. The descriptions are produced in two languages (English and French), and they are rated in terms of communicative adequacy. This should allow students not only to learn how to produce correct sentences, but also help them to realize which one of them is, communicatively speaking, the most adequate form.

1 Introduction

We live in a world surrounded by objects that we can see, touch, smell or use, and to which we can refer via our body (eyes, gestures) or via language. In order to allow others to 'see' the objects we have in mind, we can point at them or describe them in linguistic terms. Being one of the first things children learn (Matthews et al., 1997) generation of referring expressions (GRE) seems to be a simple task, yet it can be quite complex, and there are various reasons for this.

While the objects we talk about hardly ever change during our conversation, the

linguistic means used for describing them do. In fact, they vary considerably. One may use a basic-level word (dog) or a more specific term (spaniel), or one may switch from a relational description (the dog *on* the lawn) to the object's role (shepherd dog). One may also use a proper noun (Fido) or simply a pointer (pronoun), etc. Imagine a scene containing a 'dog' and a 'cat', the goal being to talk about the 'dog'.

entity	type	size	color
e_1	cat	small	white
e_2	dog	big	black

In this particular case, only the type (dog) needs to be mentionned. Actually, in theory any of the following expressions would do: (1) it, (2) the Doberman, (3) the dog. All forms are correct, but not all of them suit equally well the situation. Expression (1) is underspecified and appropriate only under very specific circumstances: high saliency, second mention, etc. (2) is overspecified, as it provides more information than needed for identifying the intended object. The third example, is a minimal description. Using a basic level term it suits most situations.

It should be noted though that minimal descriptions are not always used. People often produce descriptions more specific than needed [7]. Actually, this extra information may be warranted. For example, in the scene described above, it would be more appropriate to use the word 'Doberman' (specific term) than the more general term 'dog'. While the latter allows recognition of the indended referent, it does not convey the possibly important information that the object referred to is potentially dangerous.

As we can see, speakers have quite a few options. Yet, if this diversity of (conceptual and linguistic) means offers the speaker much freedom, it also puts pressure on him. He has to choose, and has to know when to use what resource. Since different forms yield different effects, the speaker has to learn not only *how* to refer to a given object —i.e. what information to provide to single out an object among others, and how to pack this information into words— but also *when* to use each specific resource: noun vs. pronoun; specific term vs. more general term. This implies that the speaker becomes sensitive to the listener —What does she know? What is she interested in? What is currently on her mind?— or otherwise he is likely to produce misleading cues. Clearly, linguistic knowledge is insufficient, we also need knowledge concerning language use (pragmatic) and peoples' information processing habits and needs. This kind of information will not be found in grammars or in textbooks, it is generally acquired by observing language in use.

While there are different referential acts, we will confine ourselves here only to one of them: descriptions of object mentionned for the first time. We will show how an object description may vary as a function of a set of alternatives.

2 The Importance of Context for Message Encoding

Olson [6] has shown that object descriptions are context sensitive, that is, the way an object is described depends on the objects from which it must be discriminated. To check his intuition, he carried out a small experiment composed of a set of simple geometrical elements. Keeping the same reference object, he asked his subjects to

verbalize the scene in varied contexts. The results showed that the same 'flat, round, white object' was described quite differently as "the white one, the round one, or the round white one" depending on the similarities between the *target object* and the elements of the *alternatives* (i.e. contrast set or surroundings elements) from which it needed to be discriminated. As one can see in Figure 1, two features are sufficient to ensure unambiguous reference: 'color', 'shape' or their combination.

Object of thought	Alternatives	Linguistic expression	Discrimination factor
● (white dot on black)	●	the white one	COLOR
●	□	the round one	SHAPE
●	● □ ■	the round white one	COLOR + SHAPE

Fig. 1. Relationship between some *intended object*, a context and the *linguistic form*.

We will show in the remainder of this paper how the notion of microworld —a scene whose description varies as a function of the changes of a set of parameters: the object to be described, the viewpoint, the context, i.e. the other objects composing the scene, objects from which it must be discriminated,— could be used to support foreign language learning, in particular, the acquisition of language use (communicative competency). Just as linguistic knowledge (the knowledge of words and rules) does not guarantee fluency (or, fluent speaking), does the skill of speaking guarantee successful communication. This latter requires not only the ability to convey what one wants to say, but also the ability to choose the right resources (content and linguistic form) at the right moment, that is, in agreement with the circumstances (context, goal), so that the listener can decode properly the message and its communicative goal. In other words, it is not enough to dump some message on the receiver to make communication happen. The speaker has to learn to move from his egocentric point of view —(What is his concern? What does he want to say?),— to the listener's position, to be able to see things also from her perspective. What does s/he know? What is on her mind? What does she believe in or care for? We will try to achieve this for a very small domain of language: the province of reference.

If mastering a language requires the learning of a skill (speaking, listening, etc.), i.e. procedural knowledge, it also requires the learning of a set of rules (declarative knowledge). Of course, this latter can be achieved in various ways, via explicit learning or by observation (our approach), etc. While this latter is the natural approach, it has a shortcoming: it is not systematic, hence we may have to wait a long time before encountering again a similar situation in which a given expression appeared for the first time. This is where a micro-world approach might be useful.

While systematic variations are certainly not part of nature's manifestations, certain kind of rules can probably best be observed (and possibly learned) in closely

controlled experiments/settings. Put differently, a lot of knowledge is acquired on the basis of regular changes or co-variation between an input and an output. Since this regularity or systematicity does not occur in nature, while it would be good for learning, one should allow for it. In addition one should visualize the determining factors (the links between an input and output) and allow the user to control the inputs.

We now describe the principles and the implementation of a web-based application, called WebREG, to help a speaker learn how to generate expressions that refer properly, i.e. unambiguously and, given the circumstances (context), adequately to concrete entities occuring in a given scene.

3 A Solution: Production of Distinguishing Descriptions

According to [1] a referring expression (RE) designates an entity in the real (or imaginary) world. This work deals with such expressions and complies with the following characteristics [2]:

- They are expressed as noun phrases rather than pronouns or other linguistic means;
- They refer to physical objects rather than to abstract entities;
- They are meant to allow the hearer *identify* an object rather than satisfy any other communicative goal.

So, following Dale & Reiter, we try to achieve a referential communicative goal by generating a distinguishing description of the target entity. This description applies to the target but not to any other object present in the scene and seen by the hearer (context set). This set, minus the target, is called the 'contrast set' or 'potential distractors'.

To distinguish the target from the potential competitors, we use features, i.e. properties (size, color, …) and relational terms (left of, in front, …). The object type (chair, sofa, fan, …) is special and mandatory for generation because it serves as the head of the noun phrase. Properties and relations are represented as attribute-value pairs. On the language side they are realized as adjectives or prepositional phrases modifying the noun phrase. For example, the object 'chair' with the property (color: red) and the relation (right of, desk) might be realized as the 'red chair to the right of the desk'. The goal of referring expression generation is thus to select a set of pairs that distinguish the target from the rest (potential alternatives, i.e. competitors).

In order to determine which RE to block, we interpret the Gricean Maxims [5] in this context as follows:

- **Quality:** a RE must be a precise description of the target;
- **Quantity:** a RE should contain enough information to allow the hearer to identify the target, but not more;
- **Relevance:** a RE should only mention features necessary to allow the distinction of the target from the contrast set;
- **Manner:** a RE should be as short as possible.

The interpretation of these maxims depends on the application goal and the context and will be discussed after the presentation of our application in the next section.

3.1 A Web Application for Learning Referring Expressions

WebREG is a web-based application that presents several types of bilingual (French and English) RE generation activities. It can be accessed on-line at:

http://www.iro.umontreal.ca/~lapalme/WebREG

The user interface (see Figure 2) shows a 2D graphic scene composed of a small set of entities represented as pictures. Entities have properties and entities can be related to each other. The entities, properties and their pictorial representation have been taken from the 'TUNA referring expression corpus' [6]. Depending on the activity, the scene's configuration, —i.e. the choice of the entity to be expressed, the set of competing entities, their position and properties— are chosen by the program or the learner.
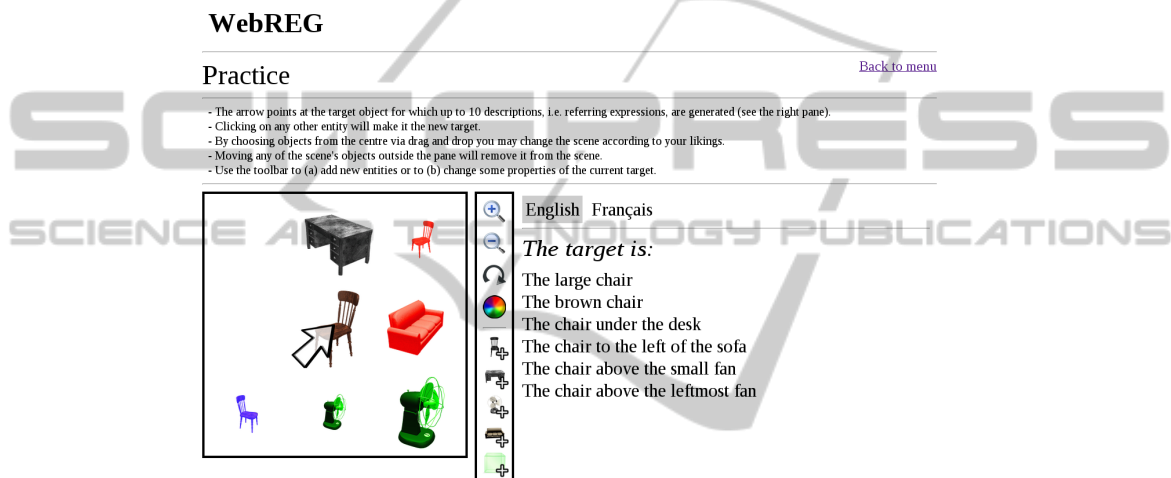

WebREG

Practice   Back to menu

- The arrow points at the target object for which up to 10 descriptions, i.e. referring expressions, are generated (see the right pane).
- Clicking on any other entity will make it the new target.
- By choosing objects from the centre via drag and drop you may change the scene according to your likings.
- Moving any of the scene's objects outside the pane will remove it from the scene.
- Use the toolbar to (a) add new entities or to (b) change some properties of the current target.

English Français

The target is:

The large chair
The brown chair
The chair under the desk
The chair to the left of the sofa
The chair above the small fan
The chair above the leftmost fan

Fig. 2. Training mode of **WebREG**: the scene is on the left, the centre contains the toolbox and to the right is the list of proposed REs for the current target pointed out by the white arrow.

WebREG can generate in French or English one or several REs for a given target . The user interface is currently only in English.

3.2 Entities, Properties and Relations

The TUNA corpus contains entities for furniture: chair, desk, fan and sofa. We added in our application a box to allow for the use of such objects. In addition we used some of the properties of TUNA like: *size* (small or large), *color* (red, black, brown, gray, green or blue) and *orientation* (from the front / back, turned to the left and to the right). Note, that not all properties apply to all objects. Boxes are always big, from the front and colored red, green or blue.

Entities are placed on a 2D plane and are referred to with respect to the other objects composing the scene. To allow for this we use the following properties and relations:

Properties:

- **absolute positioning:** top left/right corner, bottom left/right corner;
- **relative positioning:** bottommost, topmost, leftmost, rightmost.

Relations:

- **positioning:** left of, right of, over, under;
- **containment:** in, contains.

3.3 Activities

We vary the learning methods in order to motivate the learner to create a RE. WebREG offers currently three types of activities. They all focus on insightful learning, i.e. competency.

3.3.1 Practice (*How to characterize or express an object ?* Figure 2)

The learner or the program create a scene by choosing some objects, placing them on a 2D plane and deciding on a target. The machine will then produce ten REs in decreasing order of relevance, the order being controlled via an algorithm described in section 4 (Reference generation). It should be noted though, that learning the principles underlying the generation of referring expressions is much harder than, say, learning the rules for generating plural nouns in English. While a single change at the input (one vs. several) may yield various changes at the output (1 car/horse vs. 2 cars/hors<u>es</u>), neither of them has an impact on the communicative level. This is quite different from the case of referring expressions where the change of a grammatical resource (say, *pronoun* rather than a *definite noun phrases*) may considerably affect the scope, i.e. the interpretation of meaning and the ease of processing.

To check the validity of this last statement let us see what happens if you vary the linguistic means [in/*definite description* ("a/the + N"), *pronouns* (its), ...] used to refer to some concepts. Suppose the concepts were POPULATION and PLACE, that they are part of a message called B, i.e. —[LEAVE (POPULATION, PLACE)],— which is preceded by A and followed by C. In this case we could have any of the following realisations (A) X-town was a blooming city. Yet, when hooligans started to invade the place, (B), [i.e. any of the occurrences (a-e) here below]. (C) The place was not livable any more.

> (a) the <u>*place*</u> was abandoned by (its/the population)/them.
> (b) the <u>*city*</u> was abandoned by its/the population.
> (c) <u>*it*</u> was abandoned by its/the population.
> (d) its/the population abandoned the <u>*city*</u>.
> (e) its/the population abandoned <u>*it*</u>.

While all the candidate sentences in (a - e) are basically well-formed, each one has a specific effect, and not all of them are equally felicitous. Some are ruled out by virtue of poor textual choices —(e.g., in (a) "*the place*" is suboptimal, since it immediately repeats a word),— others because of highlighting the wrong element, or because of wrong assignment of the informational status (given-new) of some element (e.g., in (d) "*the city*" is marked as 'minimal' *new* information, while actually it is known, i.e. *old* information). Probably the best option here is (c) since this preserves

the given-new distribution appropriately, without introducing potentially ambiguous pronouns.

3.3.2 Finding the Target (*Where is the target?* Figure 3)

The system generates randomly a scene and selects a target. While not highlighting the target, it displays a description of it (in this particular case: 'the fan facing to the left') inviting the user to find the object described by the RE (here, 'the small red fan at the lower left side of the scene').

Fig. 3. Target mode of **WebREG**: the learner must click on the entity she thinks the RE is referring to. In the top right corner we see the number of attempts and the percentage of correct answers.

Learners tend to make typical mistakes. This being so, one can imagine scenarios with typical confusion sets, i.e. frequently confused items are presented together as contrastive sets. This kind of strategy is frequently used in other areas of language learning like spelling or pronunciation.

3.3.3 Decide on the Correct RE (*Which referring expression is correct?* Figure 4)

The system generates randomly a scene, selects a target and shows it to the user. In addition it generates REs for all objects of the scene, inviting the user to find the RE corresponding to the target.

The user can change the language at any moment and the system will generate accordingly the RE in the chosen language. This kind of learning can be seen as a combination of the preceding two exercices : learning by observing correlations and learning based on contrast sets.

WebREG

Which Referring Expression is correct?   **Wrong**   Click to continue
Back to menu

The system has generated a scene, selected a target and produced a Referring Expression (RE) for each entity of the scene.
It is your task to find out which of these descriptions describes the target object. To show your decision click on the RE that designates the target.

English Français   #29 / 69%

The target is:

The sofa
The bottommost desk
The desk above the small fan
The desk in the top right corner
The leftmost desk
The large fan
The small fan

Fig. 4. The learner must click on the RE on the right corresponding to the target pointed at by the arrow. If ever the user made a mistake, the system displays the right answer in green (both the RE and the object), while highlighting in red the RE selected by user and the corresponding entity.

4 RE Generation from an Algorithmic Point of View

RE generation involves finding a set of property-value and relation-value pairs to designate the target unambiguously. As our scenes are composed of only a small set of entities and relations, we decided to select an algorithm that tries all possible solutions up to a certain limit to remain in line with psycholinguistic constraints and the Gricean maxims.

The algorithm is based on a branch and bound algorithm that guarantees the production of an unambiguous RE for a given target [5]. It determines a single optimal RE for a scene and a target. This strategy is perfectly fine in cases where the goal is the production of a single expression, but our problem is slightly different. Depending on the activity we want to offer the learner more than one RE. In addition we would like to have a finer control concerning the target's interpretation (possible ambiguity) and the resources used in order to refer to an object.

Hence, we modified the algorithm to keep the N-best subgraphs according to the cost function which roughly corresponds to their number of arcs. Yousfi-Monod [10] presents the changes made to the original algorithm in order to produce more *natural* sounding descriptions in the case of many similar objects or when the referring expressions tend to become too long.

The N-best subgraphs are used to generate *N* REs using a generic noun phrase pattern comprising a determiner, a noun, adjectives and prepositional phrases which are themselves composed of a preposition and a noun phrase. The graph is recursively traversed starting from the vertex corresponding to the target. The type property is converted to a noun, the others to adjectives and relations map to prepositions.

A specialized lexicon has been built for entities, properties and relations. It

contains the necessary morphological information in order to allow for adequate processing of gender and number in English and French.

When being in the practice mode, WebREG presents many realizations for the target. Currently they are shown in the order in which they are generated by the algorithm described above. However, one could well sort them according to other criteria like: number of words, or types of properties and relations, etc.

WebREG is a client-server web application. The client side is executed in the web browser via a Javascript —(we use the jQuery framework to deal with the specifics of each browser)— to modify the HTML and CSS to change the position, size and color of the objects. The Javascript code generates the scene and deals with the users' actions. It communicates the modification of the scene to the server that deals with the generation of REs, and it keeps track of the learners' progress.

5 Perspectives

This prototype opens new possibilities: the current properties are quite generic and can be applied to all entities. Adding new ones implies not only that we have to deal with them at the algorithmic level, but also at the interface level, i.e. how to make them natural, i.e. easy at the user interface. Other types of activities could be added such as:

- **Re-create the Input:** the system creates a scene, identifies a target and generates a RE. Yet, rather than showing the initial input (scene) the system shows a modified version of it asking the learner to change it to comply with the produced RE.
- **Create the Correct RE:** pointing at a component element of a scene (target object) the system presents a list of words to be used by the user (via drag and drop) to create the appropriate RE.
- **Select the Most Natural RE:** since objects can be described in many ways, the user should be allowed to select the one seeming to be the most natural one to him or her. The system could track these answers and determine human preference criteria (relative or absolute positioning, adjective vs. propositional groups, …) which could be used later on for prioritizing REs. In this case, it would be the system that learns, not the user. This would also validate the intuitions we embedded in the system and our interpretation of the Gricean maxims.

It would be interesting to allow for several targets rather than just one to parameterize the difficulty of exercises.

Of all the proposed activities, the last one is probably the most interesting one, as it gives researchers a means to collect natural data to design then their algorithms accordingly. Of course, this makes only sense if the sentences proposed by the system contain natural sounding forms. Still, this looks like a very promising strategy, as we are still not in a position to guarantee that the chosen means (linguistic resources) allow both well-formedness and sucessfull communication. This being so, it may be wise to listen to the user asking them for feedback concerning the naturalness of a given expression produced under certain (well controlled) constraints.

6 Conclusions

This work is, to our knowledge, the first one to use the generation of REs as a task for learning a new language. As stressed at the beginning, learning a language is more than just learning to produce grammatically correct forms. As anything can be expressed via various means (lexical, grammatical, ...), one also needs to learn when to use what specific resource. And in this respect REs are an interesting application, as even native speakers do occasionally make 'mistakes', not at the linguistic level, but at the pragmatic side.

We have extensively drawn on features of the TUNA challenge [3]. For example, the selection of discriminating features is based on an algorithm that was very successful within this context. We adapted this algorithm to be able to produce various acceptable forms for a given target while being able to address in an efficient and useful way for the learner some of the inherent ambiguities of the target. It would be interesting now to confront our system with the real world and test it with language learners. To this end we will certainly revise the icons used, and more importantly, see what kind of principles could be used to produce expressions that are not only grammatically correct, but also natural, i.e. corresponding to the forms used by 'normal' people, that is, a majority.

References

1. Dale, R. (1992). Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes. MIT Press, Cambridge, MA.
2. Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 19(2):233–263.
3. Gatt, A., Belz, A. & Kow, E. (2009). The TUNA-REG challenge 2009 : Overview and evaluation results. In ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation, pages 174–182, Morristown, NJ, USA.
4. Grice, P. (1975). "Logic and conversation". In Syntax and Semantics, 3: Speech Acts, ed. P. Cole & J. Morgan. New York: Academic Press
5. Krahmer, E., van Erk, S. & Verleg, A. (2003). Graph-based generation of referring expressions. Computational Linguistics, 29(1):53–72.
6. Olson D.R. (1970). Language and Thought: Aspects of a Cognitive Theory of Semantics. Psychological Review, 77:257-273.
7. Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89-110.
8. Reiter, E.. (1990). Generating descriptions that exploit a user's domain knowledge. In Current research in natural language generation, Dale, R., Mellish, C. & M. Zock (Eds.). Academic Press Cognitive Science Series, Vol. 4. Academic Press Professional, Inc., San Diego, CA, USA, pp. 257-285.
9. van Deemter, K., van der Sluis, I. & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In INLG '06: Proceedings of the Fourth International Natural Language Generation Conference, pages 130–132, Morristown, NJ, USA.
10. Yousfi-Monod, M. (2010). WebREG: un outil pour apprendre les expressions référentielles, http://atour.iro.umontreal.ca/drupal7/sites/default/files/publis/WebREG-06-2010.pdf, internal report, RALI - Université de Montréal