university of groningen

University Medical Center Groningen

University of Groningen

An Empirically Motivated Algorithm for the Generation of Multimodal Referring Expressions

van der Sluis, Ielke

Published in:
Proceedings of the Student Research Workshop of the 39th Annual Meeting of the Association of Computational Linguistics (ACL'01)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Early version, also known as pre-print

Publication date:
2001

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
van der Sluis, I. (2001). An Empirically Motivated Algorithm for the Generation of Multimodal Referring Expressions. In *Proceedings of the Student Research Workshop of the 39th Annual Meeting of the Association of Computational Linguistics (ACL'01)* (pp. 67-72).

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 26-03-2023

An empirically motivated algorithm for the generation of multimodal referring expressions

Ielka van der Sluis
Computational Linguistics
Tilburg University
I.F.vdrSluis@kub.nl

Abstract

We describe an algorithm for generating multimodal referring expressions, based on empirical data. The main novelties are (1) a decision to point based on both the efficiency of pointing (Fitt's law) and the inefficiency of a full linguistic description, (2) the explicit tracking of the 'focus of attention', and (3) a threedimensional notion of salience incorporating linguistic, focus and inherent salience.

1 Introduction

The multimodal algorithm we discuss in this paper is based on a version of the Incremental Algorithm due to Dale and Reiter (1995), which is one of the most widely used algorithms for the generation of definite descriptions. Essentially this algorithm iterates through the list of preferred attributes, adding a property to the description for an object only if it rules out one or more of the remaining objects in the distractor set. According to Dale and Reiter, the Incremental Algorithm has at least two important properties: it is computationally attractive, because it has a polynomial complexity, and it is psychologically realistic, because it appears that humans produce distinguishing descriptions in a similar way.

There are at least two motivations for a multimodal extension. First of all, in various situations a purely linguistic description may simply be too complex, for instance because the domain contains many highly similar objects. In that case, a deictic, pointing gesture is the most efficient way of singling out the target referent. The second reason is that when looking at human communication, referring expressions which include pointing gestures are very common. Since our aim is to generate descriptions in a realistic way, it is expedient to include such pointing gestures.

The multimodal extensions to the incremental algorithm which we propose are based on empirical studies of how human speakers refer to objects in a shared work-space (Beun and Cremers, 1998), which we will summarize in section 2. The Incremental Algorithm is outlined in section 3. In section 4 we show how the empirical rules can be captured in a formal and computational manner. The main ingredients of our algorithm are the following: (1) we define a function which determines whether a pointing gesture is opportune given the current context, (2) the algorithm explicitly tracks the focus of attention, and (3) we distinguish various reasons for which a particular object might be more salient than others. We end with some concluding remarks in section 5.

2 Empirical observations on multimodal object descriptions

In this section we present four rules for referring to objects in a multimodal setting, derived from the empirical results reported by (Beun and Cremers, 1998). Beun and Cremers, who studied the way humans refer to objects, performed several experiments in which one participant had to instruct another participant (in Dutch) to make changes in a block building that was located in a shared workspace. This implied that participants

could both talk about and point to the blocks in front of them. The rules we will use for our multimodal algorithm are the following:

Rule 1 If the target object is inherently salient within the domain of conversation, use reduced information.[1]

Rule 2 If the target object is located in the current focus area use only information that distinguishes the object from other objects in the focus area.

Rule 3 Use absolute features (e.g., color) as much as possible and use relative features (e.g., size) only when necessary.

Rule 4 If an explicit relatum is needed for referring to the target object, choose as a relatum an object that is salient.[2]

Note that by ordering the properties of an object according to human preference as Dale and Reiter propose, rule 3 is already included in the algorithm. In section 4 we cover rule 2, partly with our definition of focus space, and partly with our notion of salience. The threedimensional notion of salience we present in section 4.3 also attends to rule 1 and rule 4.

3 A sketch of the Incremental Algorithm

The aim of the Incremental Algorithm (Dale and Reiter, 1995) (henceforth referred to as the D & R algorithm) is to efficiently generate a distinguishing description; a description that is applicable to the current object and not to any other object in the domain of conversation. Objects in a domain can be characterized in terms of a set of attribute-value $\langle$ A,V $\rangle$ pairs corresponding to their properties.

One of the distinguishing properties of the Dale and Reiter algorithm is its use of a list of *preferred attributes*. In this list the properties relevant for the domain are ordered according to the preference that human speakers and hearers have when talking about objects in that particular domain. (This satisfies rule 3 of Beun and Cremers, section 2)

The input of the D & R algorihm consists of the *target object* r and a *distractor set*, where r is the object to be described and where the distractor set contains all the objects in the domain except r itself. The D & R algorithm essentially iterates through the list of preferred attributes, adding a property to the description for r only if it rules out one or more of the remaining objects in the distractor set. Moreover, Dale and Reiter make the assumption that the property *type* should always be included in a referential expression even if it has no discriminating power. The D & R algorithm terminates when the distractor set is empty or when all the properties of r have been checked. In the former case the algorithm has succeeded in selecting a set of properties which distinguish r from the other objects, and on the basis of these a distinguishing description can be generated. In the latter case, the algorithm fails.

Krahmer and Theune (1999) provide a number of extensions to the basic D & R algorithm. To begin with, they introduce a notion of linguistic context. The idea is that once an object has been mentioned, it is linguistically salient and rereferring to this object can be done using a reduced, anaphoric description. Linguistic salience is modelled using a *salience weight function*,[3] according to which a salience weight is added to each object in the domain of conversation. With these additional salience weights the distractor set can be specified as the set that contains all the objects in the domain having a salience weight *higher than or equal to* the target object. A second extension which Krahmer and Theune introduced is the possibility of including relations in the D & R algorithm, in such a way that a description of the relatum of the target object is generated by a recursive call to the incremental algorithm. We take this extended version of the incremental D & R algorithm as our starting point.

[1] With "reduced information" Beun and Cremers refer to the fact that less properties are generated than would be required for a uniquely referring description.

[2] When an object is referred via its relation to another object, this latter object is called the *relatum*.

[3] The notion of salience used by Krahmer and Theune is a combination of the centering approach of (Grosz et al., 1993) and the Pragueian topic/focus ordering of (Hajičová, 1993)

4 Main ingredients of the multimodal algorithm

In this section we describe the main novelties of our multimodal extension, which are based on the empirically motivated rules discussed above. We first discuss when the algorithm may include a pointing act. Second, a notion of focus of attention is specified. Third, a threedimensional notion of salience is presented and finally we discuss the linguistic realization of the referring expressions.

4.1 When to point?

We take it that the decision to use a pointing act for distinguishing an object is codetermined by two factors: the efficiency of pointing and the inefficiency of a full, linguistic description.

We assume that the efficiency of pointing is determined by the distance to and the size of the intended referent. The trade-off between these factors has been captured in Fitts' law, the index of difficulty ID (Fitts, 1954). The index of difficulty is computed from the size of the target object and the distance measure between the object and the position of the pointing device used, in our case the speaker's hand. If this index is below a certain threshold[4], i.e., it is easy to point, the algorithm includes a pointing act in the output.

Index of Difficulty (ID)

$$\text{ID} = \log_2(\frac{2\text{D}}{\text{W}})$$

where W = Width of the object, D = Distance from pointing device to the object

The second factor that contributes to the decision to point is the inefficiency of the linguistic description. We assume that this inefficiency is proportional to the number of attributes and relations needed to generate a distinguishing description. In the algorithm the complexity of a linguistic description depends on the informative value of the properties. When the complexity of the linguistic description is above a certain threshold, the linguistic description generated so far is discarded and a pointing act is generated instead. When the target object cannot be uniquely identified by a purely linguistic description (e.g. the distractor set is not empty after iterating through the list of properties) the algorithm also includes a pointing act. Once a pointing act is included in a referring expression the target object is uniquely identified and the distractor set is emptied.

[4]The precise value of this threshold is an empirical matter, and depends on the task and the domain at hand.

4.2 Focus space

Beun and Cremers (1998) state that to describe the target object it is only necessary to distinguish it from the objects in the current focus space (rule 2). This notion of a focus space is not only psychologically plausible, but is also beneficial from a computational point of view. By defining the focus space as a subset of the objects in the whole domain, the search space of the algorithm is generally reduced. In our definition the current focus space consists of the last mentioned object r and the set of objects directly related to r. An object d is standing in a direct relation to an object r if d is the closest object to r for which that particular relation holds. To be able to take into account the surface used by the objects in the domain of discourse, we use *perceptual grouping* (Thorisson, 1994). This is a proximity score for the distance of each object in the domain to a particular object r, in relation to the maximal distance between the object r and an object in the domain. By setting a threshold to the proximity score, far away objects are excluded from the focus space of r.

In Definition 1, the focus space of an object o is formally defined as the union of the object o itself with the set of objects in the domain that are closest to o for a relation of a given *type* and which are not 'too far away' in terms of Thorisson's notion of perceptual grouping. The proximity score is defined by PS. (The threshold T should be set on an empirical basis.)

Definition 1: Focus space

$$\begin{aligned}\text{focus\_space}(o) = \{o\} \cup \{d \in D \mid & \text{ relation}(\textit{type}, o, d) \wedge \\ & \neg\exists d'(\text{relation}(\textit{type}, o, d') \wedge \\ & (\text{distance}(o, d') \leq \text{distance}(o, d))) \wedge \\ & \text{PS} \leq \text{T}\ \}\end{aligned}$$

where: $\text{PS} = \dfrac{\text{distance}(o, d)}{\max\limits_{y \in D}\ (\text{distance}(o, y))}$

Notice that the target object r need not be an element of the current focus space of o. In that case we speak of a *focus shift*. Our definition of

focus space can be illustrated with Figure 1. Assuming that the object last mentioned is the black block, the focus space contains three blocks as shown in the picture (the black block itself plus the two white ones). The grey block to the right of the black block is excluded because there is a closer block to the right of the black block. Once the algorithm has generated a referring expression for the block indicated with an $*$, the focus space needs to be updated. The updated focus space contains $*$ and the set of objects that are directly related to $*$ which would be the black block and the grey block, in the sense that there is no object closer to $*$ standing in the same relation to $*$. However using the proximity score bound to a certain threshold, the grey block can be excluded from the set of directly related objects, because it is located too far away from $*$. The 2nd, updated focus space contains $*$ and the black block.

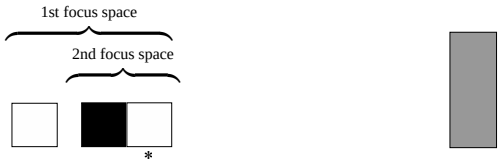

Figure 1: "the white block to the right of the black one"

4.3 A three dimensional notion of salience

Beun and Cremers (1998) have shown that various notions of salience, in addition to linguistics salience, play a role in the generation of referring expressions. In particular, objects can be inherently salient and/or they can be salient because they are in the current focus space. Consequently, in this paper we define a three dimensional notion of salience. More precisely, each object in the domain receives three salience weights: LC indicating whether or not the object is salient in the linguistic context, IS indicating whether the object is inherently salient and FS indicating the focus space salience of the object. The total salience weight of an object is defined as the sum of these three separate salience weights.

Some forms of salience are more important than others. We assume that linguistic context salience is primary, an object which was just described is more salient than an object which is in the current focus space but has not itself been mentioned. In line with the results of Beun and Cremers we take it that an object which is in focus is more salient than an object which is inherently salient but falls outside of the current focus space (rule 1,2).

Following (Krahmer and Theune, 1999) we restrict the distractor set to objects which are at least as salient as the intended referent. This implies that when the intended referent is somehow salient, the search space is reduced. In this way, generally fewer properties are required to empty the distractor set. In our algorithm, the notion of salience also plays an important role for the selection of relatums. In particular, following rule 4 Beun and Cremers, when a relatum is needed to describe an object the algorithm selects the most salient object that enables it to generate a distinguishing expression.

Definition 2 calculates the salience weight of each object d in a state s_i by taking the sum of the three kinds of salience associated with d. In the initial state s_0 (the beginning of the discourse) no object has been described and consequently we assume that there is no focus space as well.

Within the algorithm presented here, LC is modelled as it is done by Krahmer and Theune (1999), who determine linguistic salience on the basis of the ranking of forward looking centers according to centering theory (Grosz et al., 1993) augmented with a notion of recency based on (Hajičová, 1993). LC-salience weights range from 0 (complete non-salience) to 10 (maximum salience). In Centering Theory, $C_f(U_i)$ is the ordering of the forward looking centers of U_i (the sentence uttered at time i) This ordering is such that the syntactic subject of U_i is the most salient object (mapped to salience weight 10) followed by the indirect object (mapped to 9) and the other objects (mapped to 8). Focus space salience (FS) is easily determined given definition 1, where an object has an FS-salience weight of 2 iff it is part of the current focus space. The FS-salience weight 2 is assigned to every object d in the focus space of object o, where o is the most recently described object (or, slightly more general, the object with the highest LC-salience). Finally, inherent salience (I) is defined here in terms of a strong criterion, stating that an object

is inherently salient only if for some attribute it has a particular value V_1 while the other objects in the domain all have a different value V_2 for that particular attribute. If an object is inherently salient, it has a constant I-salience weight of 1.

Definition 2: Three Kinds of Salience

For each object $d \in D$, the salience weight of d in state s_i is salience_weight$(d, s_i) = \mathrm{I}(d, s_i) + \mathrm{LC}(d, s_i) + \mathrm{FS}(d, s_i)$ where:

- *Linguistic Context Salience*

$$\mathrm{LC}(d, s_0) = 0$$

$$\mathrm{LC}(d, s_{i+1}) = \begin{cases} \mathsf{rank}(d, C_f(U_i)) & \text{if } d \in C_f(U_i) \\ \mathsf{max}(0, sw_i(d) - 1) & \text{otherwise} \end{cases}$$

- *Focusspace Salience*

$$\mathrm{FS}(d, s_0) = 0$$

$$\mathrm{FS}(d, s_{i+1}) = \begin{cases} 2 & \text{if } d \in \text{focus\_space}(o) \wedge o = \\ & \text{first\_element\_of}(C_f(U_i)) \\ 0 & \text{otherwise} \end{cases}$$

- *Inherent Salience*

$$\mathrm{I}(d, s_i) = \begin{cases} 1 & \text{if object } d \text{ is inherently salient} \\ 0 & \text{otherwise} \end{cases}$$

4.4 Linguistic Realization

In this subsection we discuss a number of aspects of the actual linguistic realization module. To determine the form of the multimodal referring expressions we inspected the corpus collected by Beun and Cremers. Note that we can determine the list of preferred attributes for the block domain used in their experiments by simply counting occurences of properties. Table 4.4 contains the distribution of the attributes used in 141 referring expressions for the block domain of Beun and Cremers.

attribute	+ *Point*	− *Point*	total
Color	38	42	80
Location	4	19	23
Shape	3	10	13
Type	5	8	13
None	11	1	12

Table 1: Selected attributes as a function of pointing acts, derived from the corpus of Beun and Cremers.

It is clear that color is by far the most preferred attribute in this domain. The attribute *type* is the least preferred attribute in this domain. This is not surprising since all objects in this domain are of the block type, which makes this is very uninformative property. However, the D & R algorithm stipulates that *type* should always be included in the final description, even if it is not discriminating. We conjecture that it should not be the type attribute which is always included, but rather the most preferred attribute for a particular domain. This is supported by the finding that in the vast majority of the cases where a pointing act is included (which uniquely identifies the target object), also the color attribute is realized (even though it is, strictly speaking, redundant).

We also inspected the corpus of Beun and Cremers with respect to the choice of determiner. In this (Dutch) corpus, demonstrative determiners are preferred over articles. It has been claimed (Piwek and Cremers, 1996) that Dutch proximate demonstratives (deze/dit; 'this') are preferred when referring to objects which are relatively hard to access. The use of distal demonstratives (die/dat; 'that') is equally distributed over more and less accessible referents.[5] Piwek and Cremers' notion of accessibility can be defined in terms of the current paper as follows:

Definition 3: Accessibility

$$\text{access}(r, s_i) = \begin{cases} \text{False} & \text{if } ((\mathrm{I}(r, s_i) = 0) \vee \\ & (\mathrm{LC}(r, s_i) \leq 8) \vee \\ & (\mathrm{FS}(r, s_i) = 0) \\ \text{True} & \text{otherwise} \end{cases}$$

Except for the accessibility of the object, the choice of determiner also depends on the co-occurence of a relatum and pointing acts. In the data from Beun and Cremers' corpus, proximate demonstratives are always accompanied by a pointing act and are never used in combination with a relatum. For the relatum itself a definite article is used. Piwek and Cremers (1996) conclude that distal demonstratives are preferred without a pointing act in case they are used to refer to accessible entities.

In sum, the resulting algorithm generates a variety of 'multimodal' NPs including the most preferred attribute (for the block domain the property *color*), instead of the *type* of the object.

[5] We are aware of the fact that there are certain differences between English and Dutch where determiners are concerned. Our algorithm, primarily based on Dutch data, formalizes the findings of Piwek and Cremers for Dutch demonstratives. For the generation of English referring expressions, some minor changes in the selection of determiners are required.

5 Remarks

In this paper we have presented an algorithm for the generation of referring expressions in a multimodal setting. The main ingredients of the algorithm are the following: (1) the inclusion of deictic pointing gestures in referring expressions, (2) the algorithm explicitly keeps track of the current focus space and (3) each object in the domain is assigned a threedimensional salience weight. The resulting algorithm is capable of generating a variety of NPs depending on the accessibility of the object, the inclusion of a relatum and/or the inclusion of a pointing gesture. For more details we refer to (van der Sluis and Krahmer, 2001).

As noted in section 3, the D & R algorithm has two attractive properties: it is computationally attractive and psychologically realistic. To what extent has our proposed algorithm inherited these properties? Given that our extensions are motivated by the empirical work of Cremers and Beun, our algorithm can be claimed to model the way humans refer to objects in a multi-modal setting. In this respect, the multi-modal algorithm presented here is arguably as psychologically realistic as the D & R algorithm. It also captures more of the variety found in human object references than the Incremental Algorithm does.[6] Computationally, the D & R algorithm is so efficient because there is no possibility of backtracking. Unfortunately, as soon as we include relations, this property can not be kept. Because there is no guarantee that the right relatum is always immediately selected, the algorithm is NP complete in the worst case. However, two factors are noteworthy here. First, the use of salience guides the search for a relatum, and offers a substantial reduction of the search space. Second, we define an upper bound to the number of properties and relations that can be included in the description. When this maximum value is reached, a pointing act is used instead of a full linguistical expression. Fortunately, as soon as such an upper-bound is defined, we regain polynomial complexity (see e.g., (van Deemter, 2001)).

[6] Unfortunately, it is difficult to put figures to this claim. Obviously, the current algorithm covers most of the data of the described by Beun and Cremers (our 'training data'), but this does not warrant any serious conclusions, given that (a) it is generally bad practice to test on training data and (b) the corpus is relatively small.

Acknowledgements This paper is based on joint work with Emiel Krahmer. Thanks are due to Paul Piwek and Harry Bunt for discussion.

References

R. J. Beun and A. H. M. Cremers. 1998. Object reference in a shared domain of conversation. *Pragmatics & Cognition*, 6(1/2).

R. Dale and E. Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

P. M. Fitts. 1954. The information capacity of the human motor system in controlling amplitude of movement. *Journal of Experimental Psychology*, 47:381–391.

B. Grosz, A. Joshi, and S. Weinstein. 1993. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

E. Hajičová. 1993. Issues of sentence structure and discourse patterns. In *Theoretical and Computational Linguistics*, volume 2. Charles University Prague.

E. Krahmer and M. Theune. 1999. Efficient generation of descriptions in context. In *Proceedings of the ESSLLI workshop on the generation of nominals*. Utrecht, The Netherlands.

P. L. A. Piwek and A. H. M. Cremers. 1996. Dutch and english demonstratives: A comparison. *Language Sciences*, 18(3-4):835–851.

K. R. Thorisson. 1994. Simulated perceptual grouping: An application to human computer interaction. In *Proceedings of the Annual Conference of Cognitive Science Society*, volume 16, pages 876–881, Atlanta. Annual Conference of Cognitive Science Society.

K. van Deemter. 2001. Generating referring expressions: Beyond the incremental algorithm. In *Proceedings of the International Workshop on Computational Semantics*, volume 4, Tilburg, The Netherlands.

I. van der Sluis and E. Krahmer. 2001. Generating referring expressions in a multimodal context. In *Proceedings 11th CLIN*.