Natural Language Engineering **1** (1): 000–000. Printed in the United Kingdom

Influence of Personal Choices on Lexical Variability in Referring Expressions

Raquel Hervás, Javier Arroyo, Virginia Francisco,
Federico Peinado and Pablo Gervás
Departamento de Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid, 28040, Madrid, Spain

(*Received 13 May 2015*)

Abstract

Variability is inherent in human language as different people make different choices when facing the same communicative act. In Natural Language Processing, variability is a challenge. It hinders some tasks such as evaluation of generated expressions, while it constitutes an interesting resource to achieve naturalness and to avoid repetitiveness.

In this work, we present a methodological approach to study the influence of lexical variability. We apply this approach to TUNA, a corpus of referring expression lexicalizations, in order to study the use of different lexical choices. Firstly we re-annotate the TUNA corpus with new information about lexicalization, and then we analyze this re-annotation to study how people lexicalize referring expressions. The results show that people tend to be consistent when generating referring expressions. But at the same time, different people also share certain preferences.

1 Introduction

People express themselves using natural language in very different ways. As Biber (1995) reminds us, variability is inherent in human language: a single speaker will use different linguistic forms in different situations, and different speakers will say the same thing in different ways. This fact challenges both the development of NLG algorithms and their evaluation (Dale and Viethen 2010), because we cannot judge an algorithm by comparing its results against a unique gold-standard answer, as usually there are many valid answers. Moreover, algorithms could produce language utterances that are perfectly correct, but are penalized or deemed as incorrect because they have not been considered as a possible solution.

Given that naturalness is one of the objectives of NLG, it is reasonable to consider that language variability is a desirable goal in the generation of natural language. There are many fields where a NLG system must deal with the generation of text using different styles or choices. In the field of narrative generation, for example, it is quite common to generate the textual rendering of a story where many characters are involved. If there are conversations between these characters, probably they would "speak" using different vocabulary or language depending on their role in

the story. This issue is more important in interactive fields like the generation of text for interactive narratives or videogames, where characters interact through language, and each of them is expected to have a different personality and probably a different way of expression. In other kind of applications where the user has to interact with a computer, like for example dialog systems, it is useful that the system can accommodate its vocabulary and means of expression to the user. A style-based approach to the generation of text could then be adopted to use different vocabulary and expressions depending on the type of user or his expertise.

An interesting precedent in the field of language variability is the work by Hervás, Francisco and Gervás (2013). They demonstrated that the performance of a lexicalization algorithm that tries to imitate human-generated referring expressions improves greatly by modelling the particular lexical choices of each person. However, their solution is not easily applicable to similar problems because it requires having a corpus of expressions or texts generated by the very same person, which can be difficult in many domains.

In the present work we advance these ideas by studying the influence of lexical variation not only for a specific person, but for a group of people. More precisely, we are interested in determining how the utterances in a given corpus can be clustered into sets of similar lexical choices, and how these sets correspond to a single person or a group of people. This process helps to gain insight into the influence of people on lexical variation and to answer questions such as:

- Do these sets of lexical choices correspond to the choices of a single person?
- Can lexical choices be shared by a group of people?
- Is a single author loyal to his personal choices in all his texts?
- Are some lexical choices more popular than others?

We have focused on the task of Referring Expression Generation (REG). Referring expressions are only a small part of natural language, but we will see that variability issues are also involved in their generation and lexicalization. People do not compose referring expressions in a single way. For example, when referring to a grey sofa, many valid lexicalizations of the same conceptual content could be generated: *the grey sofa*, *a grey settee*, *grey couch*, *the sofa that is grey*, etc. Supposing that each of them is unambiguously distinguishing the intended referent, what makes different people choose one over another? This kind of decision is governed partly by the situation and context in which the communicative act is taking place, but also by many other factors. The diversity of the possible answers is huge even in this very restricted context.

The main goal of this work is therefore to study lexical variation and if it is affected by personal choices. We do not imply that personal preferences are the only thing that influence the lexicalization of a referring expression. However, we have observed that when different people have to lexicalize similar information they sometimes use different vocabulary and expressions. In order to explore this idea we will use a well-known corpus of referring expressions: the TUNA corpus, by Gatt, van der Sluis and van Deemter (2007). Taking a closer look at this corpus, we can find many different lexicalizations for the same conceptual information. If the

attributes expressed in a referring expression do not determine the lexicalization, there must be other aspects that influence it. The TUNA corpus is suitable to study lexical variability, because the context and the content to be expressed in the referring expressions were quite restricted in the experiment. We consider that lexical choices of different authors were the cause of this variability, including some basic syntactic choices taken during the lexicalization process. In order to study the influence of authors' choices on lexical variation in the lexicalization of referring expressions we propose a methodological approach in two steps:

1. Annotation of new lexical information. Addition of a new layer of information about the lexicalization of the referring expressions in the TUNA corpus.
2. Analysis of lexical variability. Determination of sets of similar lexical choices in the original corpus by means of an exploratory analysis of the re-annotated corpus. For this work we have used a clustering approach in order to find groupings of referring expressions that are homogeneous in terms of lexical choices.

The rest of the paper is structured as follows. Section 2 outlines some related previous work. Section 3 presents a brief outline of the TUNA corpus and a study of its heterogeneity. Section 4 presents the process of annotating the TUNA corpus with additional lexical information. Section 5 shows the analysis of the corpus in order to determine sets of similar lexical choices. Finally, Sections 6 and 7 outline the discussion, conclusions and future work.

2 Related Work

The general process of text generation takes place in several stages, during which a conceptual input is progressively refined by adding information that will shape the final text (Reiter and Dale 2000). During the initial stages the entities and messages that will appear in the final content are decided and these messages are organized into a specific order and structure (*content planning*), and particular ways of describing each entity where it appears in the discourse plan are selected (*referring expression generation*). This results in a version of the discourse plan where the contents, the structure of the discourse, and the level of detail of each entity are already fixed. The *lexicalization* stage that follows decides which specific words and phrases should be chosen to express the domain entities and relations which appear in the messages. A final stage of *surface realization* assembles all the relevant pieces into linguistically and typographically correct text.

Given that naturalness is one of the objectives of NLG, it is reasonable to consider the influence of language variability when implementing an NLG algorithm. However, this idea has only started to garner attention in recent years.

2.1 Language Variability Considerations in Natural Language Generation

Power, Scott and Bouayad-Agha (2003) describe an algorithm for generating a variety of texts that express the same content by considering a set of linguistic

constraints. These constraints are parameters like paragraph or sentence length, frequency of connectives, pronouns or punctuation marks, use of technical terms, etc. This work illustrates how to generate different texts that are equally correct for a specific communicative goal.

Reiter, Sripada, Hunter, Yu, Davy (2005) present a corpus-based analysis of how humans write weather forecasts. Their results show that the choice of words used to describe particular data depends on, among other things, the preferences of individual authors. They argue that the generation of texts by taking into account personalization would allow readers to get texts that are optimized to their personal style, thus making these texts more understandable for people with limited literacy skills or under some kind of stress (for example when receiving medical information about operations or medication).

Paiva and Evans (2005) present a natural language generation system whose behavior is controlled by correlating internal decisions with the surface stylistic characteristics expected in the resulting outputs. This approach is exemplified by a system that generates short medical information texts with stylistic variation in the sense of Biber (1988). Due to their implementation the generator decisions and the stylistic characteristics of the texts are quite independent of each other.

Mairesse and Walker (2011) present a parameterizable statistical generator whose parameters are based on psychological findings about the linguistic reflexes of personality (extroversion, emotional stability, agreeableness, conscientiousness, and openness to experience). The system is trained on personality-annotated data to perform the generation decisions required to create a text aimed at a particular personality: given a target personality, the system estimates values for each of its parameters and creates the final text according to them. Although it is a very powerful system, the definition of personalities is based on psychological aspects not related to language.

As diverse as they are, all these works seem to agree about the usefulness of allowing different possibilities during the NLG process in order to produce more natural texts. These studies also build on the idea that different people use language in their own different ways and that this diversity could be useful to develop more human-like NLG systems. In this paper, we try to contribute to these relatively unexplored path with a systematic approach.

2.2 Language Variability Considerations in Referring Expression Generation

REG is one of the most studied problems in the NLG process. Many different solutions have been proposed for this task, each one taking into account different considerations and approaches. A detailed survey on the topic can be found in Krahmer and van Deemter (2012). Although some systems may differ, REG is usually considered to be a two-step procedure. The first step, usually called *attribute selection*, is to choose the conceptual information to be expressed. This information should be enough to distinguish the entity being referred from any other distracting elements. Then, it is necessary to decide how that information will be expressed in

the text. This step requires selecting which words or expressions are more suitable for each part of the expression. This is mostly a task for the *lexicalization* stage which usually happens later in the generation process, although it is closely related to the generation of references.

Although different REG systems have tried to model the "right way" of performing this task, it is clear that there are many valid approaches (Dale and Viethen 2010). For example, each person has a different way of speaking or writing, and this could result in several possible referring expressions that are equally good for the same intended referent. This idea is gaining strength in the REG community.

Several works (Dale and Viethen 2009; Dale and Viethen 2010; Viethen and Dale 2010) have focused on the attribute selection part of the process. They study two corpora of human-produced referring expressions in visual scenes and find that individual author choices play a very important role in creating the semantic content of referring expressions in identification tasks.

Bohnet propose different algorithms (Bohnet 2008; Bohnet 2009) for attribute selection and lexicalization of references. All of them based on the individual profiles of each of the participants in the TUNA experiment. For the attribute selection, the system tries to create the set of attributes for an intended referring expression depending on the person who created the reference for the same example during the experiment. Taking into account other references made by a certain person, it tries to approximate the attribute selection shown by this person in other expressions. From the point of view of lexicalization, different models of vocabulary and syntactic expressions are created for different people. Information about frequent choices in the use of determiners, favorite words for values and syntactic preferences are stored.

Di Fabbrizio, Stent and Bangalore (2008) try also to take into account differences between speakers, along with the relation between semantic and word order information. For the attribute selection task they study the corpus in order to determine what were the most frequent usages of attributes for each of the participants in the creation of the corpus. From the point of view of lexicalization, they consider the different situations in the corpus as templates, but in such a way that if a set of attributes not present in the corpus is being lexicalized, the system fails because of lack of a suitable template.

Hervás *et al.* (2013) carried out an experiment that shows how the performance of a lexicalization algorithm that tries to imitate human-generated referring expressions improves greatly by modelling the particular choices of a specific person. A referring expression lexicalizer was tested following two different approaches: one that considered the lexical choices of the person who had originally generated a referring expression in the corpus, and another that did not. For the approach that does not take into account personal choices, the training set was composed by all the examples in the corpus at the same time. In the second approach the idea was to reproduce the lexicalization choices of different people so the corpus was divided in several training sets, one for each individual author.

All these approaches acknowledge the importance of considering personal choices in the generation of referring expressions and build on the idea that each person tends to be consistent with his own set of choices when he is expressing through

language. However, they miss the broader idea of sets of choices that can be shared by different people, and idea that could improve the applicability of this kind of systems.

3 Presenting and Analyzing the TUNA Corpus

In this section we present a brief description of the TUNA corpus and a study of the heterogeneity of lexicalizations within it. We have centered on lexicalization instead of attribute selection because annotations in the TUNA corpus are mainly oriented to conceptual attributes, and this information allowed us to study how many variations could be found in the lexical items chosen to express the same attributes in different referring expressions.

3.1 The TUNA Corpus

One of the most important projects in the field of Referring Expression Generation is the TUNA project (van Deemter *et al.* 2006) and its utility for evaluation as shown in Gatt *et al.* (2007). Under this project a corpus of referring expressions for visual entities in the domains of people and furniture was developed. This corpus was obtained during an experiment in which participants were asked to write textual descriptions for objects (target entities) in visual domains by typing and submitting referring expressions that distinguished them from other objects that were shown simultaneously (distractors). An experimental trial consisted of one or two targets, plus six distractors (other entities from the same domain), without any kind of additional context. Each description was later annotated with semantic information, including information about all the other objects that appeared in the scene. Each referring expression from the corpus is accompanied by the conceptual representation of the situation in which it was generated. An exhaustive description of the corpus can be found in Gatt, van der Sluis and van Deemter (2008).

For this work we have used the subset of references to singular targets which contains 398 data files (XML documents) for the furniture domain and 340 for the people domain. Each file consists of a single instance of the corpus, that is, a pair consisting of a single situation (the representation of entities and their attributes) and a referring expression that describes an entity in that situation (the target).

Figure 1 presents an example of one of these XML documents. The basic format of the instances of the corpus consists mainly of the following nodes:

- *DOMAIN*: Representation of entities in terms of their attributes.
- *STRING-DESCRIPTION*: The string describing the target referent in the domain.
- *ATTRIBUTE-SET*: The set of domain attributes included in the description. Attributes represent the characteristics of an entity by using attribute-value pairs. The possible attributes and values for both domains are shown in Table 1. Empty cells represent attributes that are not used in the domain. X-DIMENSION and Y-DIMENSION correspond to the coordinates of the referent

in a 5 (column) x 3 (row) matrix in which the objects were presented during the experiment.

- *DESCRIPTION*: The string in STRING-DESCRIPTION where the relevant substrings are annotated with attributes from the ATTRIBUTE-SET. The substrings corresponding to attributes could be single words (*big*) or phrases (*that is big*).

```
<TRIAL>
  <DOMAIN>
    [. . .]
  </DOMAIN>

  <STRING-DESCRIPTION>the smaller view of a green desk</STRING-DESCRIPTION>

  <ATTRIBUTE-SET>
    <ATTRIBUTE ID="a3" NAME="type" VALUE="desk"></ATTRIBUTE>
    <ATTRIBUTE ID="a2" NAME="colour" VALUE="green"></ATTRIBUTE>
    <ATTRIBUTE ID="a1" NAME="size" VALUE="small"></ATTRIBUTE>
  </ATTRIBUTE-SET>

  <DESCRIPTION>
    the
    <ATTRIBUTE ID="a1" NAME="size" VALUE="small">
      smaller
    </ATTRIBUTE>
    view of a
    <ATTRIBUTE ID="a2" NAME="colour" VALUE="green">
      green
    </ATTRIBUTE>
    <ATTRIBUTE ID="a3" NAME="type" VALUE="desk">
      desk
    </ATTRIBUTE>
  </DESCRIPTION>

</TRIAL>
```

Fig. 1. Example of referring expression from the TUNA corpus in the furniture domain.

There were 57 authors per domain who were identifiable via a unique identification number (ID). Each author created 7 expressions for the furniture domain and 6 for the people domain. Although this corpus is thoroughly annotated with conceptual information about the attributes used in the referring expressions, it lacks information about the lexical and syntactic form these attributes take.

3.2 Heterogeneity of the Lexicalizations of the TUNA Corpus

Hervás *et al.* (2013) analyzed the TUNA corpus in order to study the degree of heterogeneity of the referring expression lexicalizations that conform it. Lexical variation was analyzed from two different points of view: variation in the different lexicalizations for specific attributes, and variation in the different lexicalizations used by individual authors.

Table 2 shows the lexical variation found for the different attributes in both domains (number of different lexicalizations used per attribute vs. total number of mentions to this attribute). We have considered that two lexicalizations are different

Table 1. Attributes and their values in the two domains of TUNA.

Attribute	**Possible values**	
	Furniture	**People**
TYPE	{chair,sofa,desk,fan}	{person}
ORIENTATION	{front,back,left,right}	{front,left,right}
X-DIMENSION (column)	{1,2,3,4,5}	{1,2,3,4,5}
Y-DIMENSION (row)	{1,2,3}	{1,2,3}
SIZE	{large,small}	
COLOUR	{blue,red,green,grey}	
AGE		{young,old}
HASBEARD		{0 (`false`),1 (`true`)}
HAIRCOLOUR		{dark,light,other}
HASHAIR		{0 (`false`),1 (`true`)}
HASGLASSES		{0 (`false`),1 (`true`)}
HASSHIRT		{0 (`false`),1 (`true`)}
HASTIE		{0 (`false`),1 (`true`)}
HASSUIT		{0 (`false`),1 (`true`)}

if they are not exactly the same in a typical string comparison. The number of possible values for each attribute is also included in the table. Note that due to the nature of the corpus both lexical and syntactic structures used for the attributes are considered as lexicalizations. For example, the value "big" was sometimes lexicalized using a single word (e.g., *large*), and sometimes using a phrase (e.g., *that is big*). In this case, referring expression lexicalization is considered as both lexical choice and syntactic choice as they are interlaced in the TUNA corpus.

For each attribute the distribution of lexicalizations is mostly homogeneous for the different values. It is interesting to note that the attributes *orientation*, *x-dimension*, and *y-dimension* present high variation in both domains. In the people domain attributes *hasShirt*, *hasSuit*, and *hasTie* have a 100% variation, with a different lexicalization for each of their mentions in the corpus (although the number of mentions is very low). All those attributes with high lexical variation correspond to features that do not have an easy and fixed lexicalization.

Table 3 presents variation by author on average (Var. column) for both furniture and people domains. The total number of mentions for each attribute is shown for easier comparison of results. A value of 0% in an attribute means that authors were always *consistent* with themselves in that attribute, that is, every time an author used the same value for an attribute he used the same lexical expression. This kind of coherence is only perfect in attributes like *hasHair* or *hasTie* that were barely used. All authors present a degree of variation lower than 50% (on average) for all the attributes, and most of these values are between 0% and 25%.

Table 2. Lexical variation in the TUNA corpus for specific attributes (Hervás *et al.*, 2013)

Attribute	Different values	Different lexicalizations	Total mentions	% of variation
		Furniture		
orientation	4	79	127	62%
x-dim	5	59	105	56%
y-dim	3	53	132	40%
size	2	17	130	17%
type	4	22	371	6%
colour	4	17	344	5%
		People		
hasShirt	2	3	3	100%
orientation	3	9	9	100%
hasSuit	2	4	4	100%
hasTie	2	3	3	100%
hasHair	2	5	9	56%
x-dim	5	63	106	59%
y-dim	3	55	122	45%
age	2	7	21	33%
hasBeard	2	16	76	21%
hasGlasses	2	19	140	14%
hairColour	3	11	104	11%
type	1	11	284	4%

More details about this analysis can be found in Hervás *et al.* (2013).

4 Annotation Stage: Expanding the Annotation of the TUNA Corpus with Lexical Information

The first step in our approach consisted in re-annotating the TUNA corpus to add lexical information that was not present in the original corpus. Whereas the TUNA corpus contains an exhaustive annotation of the conceptual content of the referring expressions, it lacks some information about the lexical form chosen to express that content. To fill this gap we have re-annotated every sample of the corpus that refers to a singular entity, adding annotations about lexical information as an extension of the original annotation.

Our annotation procedure was to define the annotation schema, perform the

Table 3. Lexical variation in the TUNA corpus considering individual authors (Hervás *et al.*, 2013)

Furniture			**People**		
Attribute	**Var.**	**Total mentions**	**Attribute**	**Var.**	**Total mentions**
y-dim	25%	132	*hasGlasses*	37%	140
x-dim	10%	105	*hasBeard*	26%	76
size	8%	130	*hasShirt*	25%	3
orientation	5%	127	*y-dim*	23%	122
colour	4%	344	*x-dim*	15%	106
type	3%	371	*age*	13%	21
			orientation	13%	9
			type	12%	284
			hairColour	4%	104
			hasHair	0%	9
			hasSuit	0%	4
			hasTie	0%	3

annotation of each referring expression in the corpus with the additional information (this is done twice, with two different annotators), and finally adjudicate the double-annotated expressions into a gold standard. A validation of the annotation process by measuring inter-annotator agreement was also performed.

4.1 Annotation Schema

The annotation schema proposed adds a new layer of information to the TUNA corpus. This layer aims to reflect information about the different lexicalizations of the referring expressions. Therefore, all aspects that were considered useful in order to provide extra information about the way in which references are expressed were initially added to the annotation. This includes syntactic information when it is relevant in the lexicalization of the referring expression. The list that follows contains the identifiers, values and corresponding descriptions of all the new annotations that have been added. New annotations were organized in four different groups. The first group deals with general grammatical choices that appear in all referring expressions. The other three groups correspond to choices when expressing position (*x-dimension* and *y-dimension*), orientation (*orientation*), and the rest of attributes. Position and orientation have been chosen separately because both present more than a 40% of variation and a significant number of mentions (see Table 2 in Section 1). An exception is orientation in the people domain, but it has

been included for completion. To facilitate later data analysis, all the annotations are represented by binary values; that is, `yes` or `no` are the only possible values.

4.1.1 General Grammatical Choices

Two new elements correspond to this category, dealing with general grammatical choices that appear in all referring expressions:

- **Determiner.** It represents whether an article (called *determiner* in the TUNA corpus) was used to indicate the referent in the referring expression. Although determiners in TUNA were annotated in the attribute set, we decided to duplicate this information in the new annotations in order to have all the lexical information in the same place. Depending on the type of determiner, there are two different elements in the annotation:
  - **Definite determiner** (e.g., ***the*** *man*).
  - **Indefinite determiner** (e.g., ***a*** *chair*).

  If these two elements have `no` as value the referent has no determiner (e.g., *small green desk*).
- **Grammatical Structure.** It represents whether the referring expression has a valid grammatical structure (both predicative and nominal phrases are considered as valid). Depending on the grammatical structure of the referring expression, there are two different elements in the annotation:
  - **Predicative**. The referring expression has a verbal structure, that is, it is a predicative phrase (e.g., *the chair is gray*).
  - **Nominal**. The referring expression has a nominal structure, that is, it is a nominal phrase (e.g., *the blue chair*).

  If these two elements have `no` as value we consider that the referring expression presents no grammatical structure (e.g., *green table small straight forward*). This does not mean that the referring expression has no structure at all. For example, some referring expressions like the example above have a "list" structure. However, we have considered these cases to have no grammatical structure in our annotation.

4.1.2 Choices for the Expression of Position

The following six elements represent how the physical position of the referent is expressed. This position corresponds to the *x-dimension* and *y-dimension* attributes in the corpus.

- **Cardinal Numbers.** The position of the referent is described using cardinal numbers (e.g., *the desk in row* ***two***).
- **Cardinal Points.** The position of the referent is described using cardinal points (e.g., *man in the* ***north*** *part of the screen*).
- **Ordinal Numbers.** The position of the referent is described using ordinal numbers (e.g., *the desk in the* ***first*** *column*). The word "last" is considered as use of an ordinal number.

- **Relation to Others.** The position of the referent is described by comparison with another element (e.g., *blue fan in the middle of the screen at the top* ***with a green fan below it***).
- **Rows and Columns.** The position of the referent is described using words like 'row', 'column' or 'corner' (e.g., *second picture on* ***row*** *1* or *first picture on third* ***row***).
- **Top-Bottom and Left-Right.** The position of the referent is described with words like 'top', 'bottom', 'left', 'right' or 'middle' (e.g., *the one at the top in the* ***middle*** or *the* ***bottom*** *right most*).

Note that there are examples in the TUNA corpus that present two or more of the defined situations simultaneously. In those cases each of the previous elements is annotated accordingly.

4.1.3 Choices for the Expression of Orientation

The following five elements represent how the physical orientation of the referent is expressed. This orientation corresponds to the *orientation* attribute in the corpus. When the orientation of a part of the referent is given, it is considered also an expression of orientation (e.g. *"desk drawers facing front"*).

- **Directly.** The orientation of the referent is described directly, without using any kind of connectives or auxiliary phrases (e.g., *the* ***backview*** *desk*).
- **Gerund Phrase.** The orientation of the referent is described using a gerund phrase (e.g., *the chair* ***facing away from me***).
- **Participle Phrase.** The orientation of the referent is described using a participle phrase (e.g., *the chair* ***shown from its side***).
- **Relative Clause.** The orientation of the referent is described using a relative clause (e.g., *the chair* ***that has its back to us*** or *the chair* ***that is facing right***).
- **Using With.** The orientation of the referent is described using the prepositions 'with' or 'without' (e.g., *the chair* ***with the seat facing away*** or *the desk* ***with its back to me***).

There are examples in the TUNA corpus that present two or more of the defined situations simultaneously, in those cases all the previous elements are annotated accordingly.

4.1.4 Choices for the Expression of Non-positional Attributes

The following six elements represent how the attributes of the referent (other than position and orientation) are expressed. Only those attributes that correspond to the referent are considered.

- **Premodifier.** The attributes of the referent are described using a premodifier, without using any kind of connectives or auxiliary phrases (e.g., *the* ***bald*** *man*).

- **Gerund Phrase.** The attributes of the referent are described using a gerund phrase (e.g., *the man* ***wearing glasses***).
- **Indicative Clause.** The attributes of the referent are described using an indicative clause (e.g., *it is a fan in the middle and* ***it is small***).
- **Relative Clause.** The attributes of the referent are described using a relative clause (e.g., *the man* ***who has black hair***).
- **Using In.** The attributes of the referent are connected to the referent by the preposition 'in' (e.g., *The man* ***in suit***).
- **Using With.** The attributes of the referent are connected to the referent by the prepositions 'with' or 'without' (e.g., *the man* ***with glasses***).

Other possible values like *participle phrase* or the connector 'from' were not considered because they do not appear in the original corpus. If any example of the TUNA corpus presents two or more of the defined situations simultaneously each of the previous elements is annotated accordingly. When more than one form of expression are nested, only the main one is considered (for example *"... who is wearing"* is considered Relative Clause, ignoring the following gerund form). Although the word chosen for the referent may contain information about some of its attributes (e.g. *"the man"* contains information about the gender), that phenomenon is not considered as an attribute per se.

Table 4 summarizes the new annotation elements and shows their abbreviated names, which also correspond to the XML elements in the corpus. Figure 2 presents an example of a referring expression from the original TUNA corpus with the new annotations.

4.2 Annotation Process

Two annotators (authors of this paper) annotated the 21 features of the annotation schema for the total of 738 samples in the corpus. As the annotation advanced, the annotation schema evolved and it became necessary to re-check the annotation performed so far, sometimes by adding new elements, or even revising previous choices when a change of interpretation was made. Once the annotation was done we tested whether different annotators produced consistently similar results. If annotation data is reliable then we can infer that they have internalized a similar understanding of the annotation guidelines, and we can expect them to perform consistently under this understanding. Reliability is thus a prerequisite for demonstrating the validity of the annotation scheme (Artstein and Poesio 2008).

The simplest measure of agreement between two annotators is defined as "the percentage of judgments on which the two annotators agree when coding the same data independently" (Scott 1955). Literature has reached the conclusion that this measure must be adjusted for chance agreement in order to get figures that are comparable across studies. Simple kappa coefficient is a statistical measure or inter-annotator agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since it takes into account the agreement occurring by chance. However, Simple kappa (Randolph

Table 4. Annotation elements and their abbreviated names.

Annotation Group	Annotation Element	Abbreviated Name
General Grammar	Definite Determiner	DETERMINER-DEF
	Indefinite Determiner	DETERMINER-INDEF
	Pred. Gramm. Struct.	STRUCT-PRED
	Nom. Gramm. Struct.	STRUCT-NOM
Expression of Position	Cardinal Numbers	POS-CARDINAL-NUMBERS
	Cardinal Points	POS-CARDINAL-POINT
	Ordinal Numbers	POS-ORDINAL-NUMBERS
	Relation to Others	POS-WITH-RELATION
	Rows and Columns	POS-ROWCOLUMN
	Top-Bottom\Left-Right	POS-TOPBOTTOMLEFTRIGHT
Expression of Orientation	Directly	EXPRESSION-OF-ORIENTATION-DIRECT
	Gerund Phrase	EXPRESSION-OF-ORIENTATION-GERUND
	Participle Phrase	EXPRESSION-OF-ORIENTATION-PARTICIPLE
	Relative Clause	EXPRESSION-OF-ORIENTATION-RELATIVE
	Using With	EXPRESSION-OF-ORIENTATION-WITH
Expression of Attribute	Premodifier	EXPRESSION-OF-ATTRIBUTE-PREMOD
	Gerund Phrase	EXPRESSION-OF-ATTRIBUTE-GERUND
	Indicative Clause	EXPRESSION-OF-ATTRIBUTE-INDICATIVE
	Relative Clause	EXPRESSION-OF-ATTRIBUTE-RELATIVE
	Using In	EXPRESSION-OF-ATTRIBUTE-IN
	Using With	EXPRESSION-OF-ATTRIBUTE-WITH

2005) is known to be influenced by prevalence and bias, which can lead to the paradox of high agreement but low kappa. It also assumes that raters are restricted in how they can distribute elements across categories, which is not a typical feature of many agreement studies. We used Randolph's free-marginal multi-rater kappa (Randolph 2005; Randolph 2008) to analyze inter-evaluator agreement for each element annotated in the corpus. We selected this measure because Brennan and Prediger (1981) suggest using free-marginal kappa when annotators are not forced to assign a certain number of elements to each category. Values of kappa can range from -1.0 to 1.0, with -1.0 indicating perfect disagreement below chance, 0.0 indicating agreement equal to chance, and 1.0 indicating perfect agreement above chance. A rule of thumb is that a kappa of 0.7 or above indicates adequate inter-rater agreement.

The value of the Kappa coefficient for almost all the elements is 1.0 which indi-

```
<TRIAL>
  <DOMAIN>
    [. . .]
  </DOMAIN>

  <STRING-DESCRIPTION>the blue chair in the middle row</STRING-DESCRIPTION>

  <ATTRIBUTE-SET>
    <ATTRIBUTE ID="a3" NAME="y-dimension" VALUE="2"/>
    <ATTRIBUTE ID="a2" NAME="type" VALUE="chair"/>
    <ATTRIBUTE ID="a1" NAME="colour" VALUE="blue"/>
  </ATTRIBUTE-SET>

  <STYLE-SET DETERMINER-DEF="yes"  DETERMINER-INDEF="no"
                STRUCT-PRED="no"  STRUCT-NOM="yes"  />

  <EXPRESSION-OF-ATTRIBUTE
            EXPRESSION-OF-ATTRIBUTE-PREMOD="yes"
            EXPRESSION-OF-ATTRIBUTE-GERUND="no"
            EXPRESSION-OF-ATTRIBUTE-INDICATIVE="no"
            EXPRESSION-OF-ATTRIBUTE-RELATIVE="no"
            EXPRESSION-OF-ATTRIBUTE-IN="no"
            EXPRESSION-OF-ATTRIBUTE-WITH="no"  />

  <EXPRESSION-OF-POSITION
            POS-CARDINAL-NUMBERS="no"
            POS-CARDINAL-POINT="no"
            POS-ORDINAL-NUMBERS="no"
            POS-WITH-RELATION="no"
            POS-ROWCOLUMN="yes"
            POS-TOPBOTTOMLEFTRIGHT="yes"  />
</TRIAL>
```

Fig. 2. Example of referring expression from the TUNA corpus re-annotated with lexical information.

cates there is almost perfect agreement above chance. Elements with lower agreement are: `STRUCT-PRED` (0.9), `STRUCT-NOM` (0.8) and `POS-WITH-RELATION` (0.9), but in all cases the value is greater than 0.7 which is considered an acceptable agreement.

4.3 Adjudication of the Annotation

Once the two annotations were ready an adjudication process for the annotation was needed to obtain a gold standard. In those cases where the Kappa coefficient was 1.0 (perfect agreement above chance) the adjudication was done automatically but for those attributes with a Kappa coefficient less than 1.0 it was necessary to adjudicate the double-annotated expression into a gold standard. Adjudication of differences was conducted by discussion between the two annotators; the first author of this paper moderated these discussions and settled strong disagreements. The result of this adjudication is the whole set of singular referring expressions in the TUNA corpus annotated with information about their lexical characteristics.

4.4 Notes on the Annotation Process

In addition to the elements presented in this section, the corpus was annotated with many more elements that finally were not taken into account for this work

because they were not directly related to lexicalization. For example, the presence of new attributes not considered in the attribute set of the original corpus was also annotated (e.g., *the* ***smiling*** *man next to the* ***angry*** *man*), along with information about whether the referring expression contained information that refers to the conceptual frame of the TUNA's experiment itself instead of the object of the referring expression (e.g., *the chair at the top right* ***surrounded by a border***) or the presence of orthographic errors (e.g., *the chair in the* ***midlle***).

Several issues appeared during the annotation process concerning strange or ambiguous situations. For example, sometimes the word "frontal" (*"grey frontal table"*) can be considered an expression of orientation or an expression of position (although it is more plausibly an expression of orientation). In addition, sometimes the referent does not appear in the referring expression we are annotating. Usually in those cases the referring expression is considered as not having a valid syntactical structure (e.g. *"right middle"*).

The extended corpus is available in `http://nil.fdi.ucm.es/index.php?q=node/567`, including all the XML files in the original TUNA corpus re-annotated with the new information, along with a document describing the annotation process.

5 Analysis Stage: Influence of Authors' Choices in the Lexical Variability of the TUNA Corpus

The extended annotation of the TUNA corpus makes it possible to analyze the lexicalization of the referring expressions in the corpus. We have carried out an exploratory analysis to determine which are the different types of lexical choices that appear in the corpus. We do not impose the groups to be found but we expect them to emerge from the data. To do this we used cluster analysis, which is a method of statistical learning whose aim is to find hidden structure in the data. The goal is to divide the expressions in the corpus into several groupings, so that expressions in the same grouping are more similar to each other than to those in other groupings. The groupings will help us to identify sets of homogeneous choices and analyze whether similar choices are shared by different authors or correspond to individuals.

5.1 Cluster Analysis

Cluster analysis (Jain *et al.* 1999) is a well-known statistical technique that will serve us for the purpose of finding groupings of similar referring expressions. Let $x_1, x_2, ..., x_n$ be a set of observations (in our case referring expressions) where each observation x_i is a d-dimensional vector and d is the number of variables that characterize the observation. In our case, the referring expressions will be represented by means of d binary variables. Each binary variable will correspond with an annotation element. For example, given four annotation elements (`DETERMINER-DEF`, `DETERMINER-INDEF`, `STRUCT-PRED`, `STRUCT-NOM`), the expression *it is a red chair with the seat facing right* will be represented by the following 4-dimensional vector of binary variables $(0, 1, 1, 0)$.

The clustering algorithm used in our work is the k-means algorithm (MacQueen 1967). The k-means clustering aims to partition the set of observations into k sets or clusters $\mathbf{S} = \{S_1, ..., S_k\}$ so as to minimize the within-cluster sum of squares

$$\arg\min_{\mathbf{S}} \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_i - c_j\|^2 \tag{1}$$

where $\|x_i - c_j\|^2$ is the squared distance between observation x_i and c_j, which is the centroid of cluster S_j. The centroid of a cluster summarizes all the observations belonging to that particular cluster. In our case, the value of each variable of centroid c_j will be computed as the mode of the variable in cluster S_j, that is, the value that appears most often for that variable in the observations x_i that belong to cluster S_j, i.e. $x_i \in S_j$. Thus, the centroid will be represented also by a vector of d-dimensional vector of binary variables.

The distance between two referring expressions a and b described by d binary variables is

$$\|a - b\| = \sqrt{\sum_{i=1}^{d} (a_i - b_i)}, \tag{2}$$

where a_i and b_i are the i-th binary variable. The distance simply takes the square root of the summation of the number of binary variables present in one of the expressions and absent from the other, and vice versa. This distance was chosen because it is simple and represents in a suitable way the similarity between referring expressions. In addition, some other distances were considered in the clustering algorithm and the results were not greatly different.

The result of the clustering process is a partition of the referring expressions into k groups (or clusters). Expressions in the same group are more similar (according to the representation and the distance used) to each other than to those in other groups. The number of clusters k will be chosen after assessing the output obtained with different k values. The assessment will take into account the size and the degree of homogeneity of each cluster, and will use the domain knowledge as a guidance. In other words, a clustering result will be considered suitable if it produces homogeneous and sufficiently large clusters that can be associated to groupings of similar lexical choices that are sound in the context of the corpus.

The clustering method described will be used to carry out two analyses of the variability in the corpus. The first will only consider general grammatical choices that are shared by all referring expressions, while the second one will focus on specific grammatical choices used for referencing non-positional attributes and position.

5.2 Finding Groupings According to General Grammatical Choices

The variables considered in this case are the annotation elements `DETERMINER-DEF`, `DETERMINER-INDEF`, `STRUCT-PRED` and `STRUCT-NOM`. These variables characterize general grammatical choices that all referring expressions have: the use of determiners and their grammatical structure.

Table 5. Centroids of each cluster of the furniture domain.

	GC1	GC2	GC3	GC4	GC5	GC6
DETERMINER-DEF	0 (100%)	**1** (100%)	0 (100%)	0 (100%)	**1** (100%)	0 (100%)
DETERMINER-INDEF	0 (100%)	0 (100%)	**1** (100%)	0 (100%)	0 (100%)	**1** (100%)
STRUCT-PRED	0 (100%)	0 (100%)	0 (100%)	0 (100%)	**1** (100%)	**1** (100%)
STRUCT-NOM	**1** (100%)	**1** (96%)	**1** (96%)	0 (100%)	0 (100%)	0 (100%)

The cluster analysis tells us which preferences (that is, combination of annotation elements) are more popular in the corpus used. The null hypothesis would be that the group of referring expressions is homogeneous and that in the TUNA corpus all referring expressions were created in the same way. We carried out a cluster analysis in the furniture and people domains independently. In both domains, the number of clusters considered was $k = 6$, because on one hand, with *less clusters* we obtained a cluster with most of the references and the rest of them with only a few expressions, and in the other hand, with *more clusters* we obtained small clusters with few referring expressions that did not represent related or similar choices.

Tables 5 and 7 show the values of the centroids in the six resulting clusters for the furniture and people domains, respectively. The value of the centroid for a variable is the value that appears most often in the expressions that belong to that cluster. In parentheses the percentage of referring expressions in the cluster for whom the values of the characterizing variables are equal to those of the centroid is shown. These percentages illustrate the homogeneity of each annotation element in each cluster. A value of 100% means that all the referring expressions in the cluster have the same value for the considered value.

Tables 6 and 8 show the number of referring expressions in each cluster for the furniture (398 expressions) and people domains (340 expressions). We can see that GC1 and GC2 are the most popular ways of creating referring expressions. The second line of these tables also shows the within-cluster variability for each cluster S_j:

$$\sum_{x_i \in S_j} \|x_i - c_j\|^2 \tag{3}$$

where $\|x_i - c_j\|$ is the distance in Eq. 2, x_i is the referring expression i and c_j is the centroid of cluster S_j. Finally, the third line of these tables shows the mean within-cluster variability in each cluster. A value of 1 in this measure means that on average the individuals of the cluster disagree on one feature from the cluster centroid.

Tables 5 - 8 show that the clusters obtained in both domains are almost equal in terms of centroids, size and variability, because we can find the same sets of similar choices in terms of the grammatical structure of the referring expression and they are used in the same proportion. The result shows that in this experiment

Table 6. Number of observations, within-cluster variability and mean within-cluster variability in each cluster of the furniture domain.

	GC1	GC2	GC3	GC4	GC5	GC6
# obs.	150 (38%)	147 (37%)	53 (13%)	31 (8%)	14 (4%)	3 (1%)
total var.	0	6	2	0	0	0
mean var.	0	0.04	0.04	0	0	0

Table 7. Centroids of each cluster of the people domain.

	GC1	GC2	GC3	GC4	GC5	GC6
DETERMINER-DEF	0 (100%)	**1** (100%)	0 (100%)	0 (100%)	**1** (100%)	0 (100%)
DETERMINER-INDEF	0 (100%)	0 (100%)	**1** (100%)	0 (91%)	0 (100%)	**1** (100%)
STRUCT-PRED	0 (100%)	0 (100%)	0 (100%)	0 (100%)	**1** (100%)	**1** (100%)
STRUCT-NOM	**1** (100%)	**1** (94%)	**1** (100%)	0 (100%)	0 (100%)	0 (50%)

Table 8. Number of observations, within-cluster variability and mean within-cluster variability in each cluster of the people domain.

	GC1	GC2	GC3	GC4	GC5	GC6
# obs.	132 (39%)	126 (37%)	44 (13%)	22 (6%)	14 (4%)	2 (1%)
total var.	0	8	0	2	0	1
mean var.	0	0.06	0	0.09	0	0.5

and considering the annotation elements mentioned, the domain was not a key factor as the choices of the participants in the TUNA experiment were consistent in both domains. Clusters GC1 and GC2 represent the most popular ways of creating referring expressions, while clusters GC5 and GC6 represent the least popular ones.

Another conclusion that can be drawn from the tables is that within-cluster variability is really low. This result is not surprising giving that we have 6 clusters and that there are only 9 possible combinations of the 4 annotation elements[1]. Thus, it is easy to understand that some of the clusters present not variability because all its referring expressions have the same annotation values.

Thus, we have found a grouping structure and it is not a naive one. If there

[1] In theory the possible combinations is 2^4, but some of the annotation elements cannot be present at the same time (namely, `DETERMINER-DEF` and `DETERMINER-INDEF`, and `STRUCT-PRED` and `STRUCT-NOM`).

were no grouping structure, we would find only one group of homogeneous referring expressions. A simpler grouping structure would be 9 groups of approximately the same size (i.e. referring expressions would be uniformly distributed among the 9 possible combinations of annotation values). However, we have found a more sophisticated clustering structure that is the same in both domains. The clusters found were:

- **Cluster GC1** corresponds to expressions with nominal structure that use neither definite nor indefinite determiners. Some examples are: *plain wooden desk* and *man at the bottom with glasses*.
- **Cluster GC2** corresponds to expressions with nominal structure and where the definite determiner is used. This cluster is roughly as big as cluster GC1. Some examples of the referring expressions in the cluster are: *the red chair on the right* and *the picture top right*.
- **Cluster GC3** corresponds to expressions with nominal structure where the referent is referred to using an indefinite determiner. This cluster is similar to GC1 and GC2, but much smaller than them. This means that the use of indefinite determiners in noun phrases is much less frequent than the use of definite determiners or the use of no determiners at all. Some expressions found in cluster GC3 are: *a red chair which is facing the left hand side of the screen* and *a grey desk*.
- **Cluster GC4** corresponds to the simplest grammatical structure employed: no structure and no determiners. These expressions are composed by one word or several disjointed words. Some examples are: *2nd chair 3rd row* and *with glasses*.
- **Cluster GC5** corresponds to expressions with predicative structure and definite determiners. A closer look at the expressions in this cluster reveals that all of them use some kind of meta-information, that is, they describe the object using information of the context of the experiment (i.e., the red border surrounding the object). In all of the expressions the subject is the element being described while the predicate always refers to the red border that surrounds the object. Some examples of the referring expressions in the cluster are: *the red chair in the bottom is boxed in red* and *the guy on the bottom row is in red box*.
- **Cluster GC6** corresponds to expressions with predicative structure and indefinite determiner. The referring expressions have the structure "it is a..." and they are very unusual in the corpus. Some examples are: *it is a red chair with the seat facing right* and *it is an older man with glasses*.

If we assume that each of the above clusters represents a set of similar grammatical choices for referring to an object, it would be interesting to analyze whether each person stayed true to one set or switched between different sets during the whole experiment.

In the furniture domain, 56 authors wrote seven expressions and one person wrote just six expressions. 63% of them (36 individuals) only used one set of choices. However, 25% and 10% of the authors switched between two or three different

Table 9. Number of consistent individuals in each cluster in the furniture domain.

GC1	GC2	GC3	GC4	GC5	GC6
11 (51%)	18 (85%)	5 (66%)	0 (0%)	2 (100%)	0 (0%)

Table 10. Number of consistent individuals in each cluster in the people domain.

GC1	GC2	GC3	GC4	GC5	GC6
11 (48%)	16 (76%)	4 (55%)	0 (0%)	2 (86%)	0 (0%)

sets, respectively. Only one individual used four sets. In the people domain, 56 authors wrote six referring expressions and one person just wrote four expressions. 58% of them (33 individuals) only used one set of choices, while 33% and 9% switched between two or three different sets, respectively. No individual used four sets or more. We can again see that the results are quite similar in both domains. Therefore, although most of the grammatical variability in the corpus is produced by authors that have a personal way of expression, we also find that some authors make different choices in different expressions. These results also show that sets of these choices are not unipersonal but shared by different authors.

Tables 9 and 10 show the number of consistent individuals (persons who always create referring expressions using the same lexical choices) per cluster in the furniture and people domains, respectively. The number in parentheses is the relative number of referring expressions written by consistent individuals. It can be seen that the numbers obtained in both domains are quite similar. There is an interesting difference between GC1 and GC2. While referring expressions in cluster GC2 belong mostly to consistent authors, in cluster GC1 the percentage of referring expressions written by consistent individuals is still substantial but significantly smaller. Interestingly, nobody wrote all the referring expressions using the choices represented by cluster GC4 (the non-structured set) nor cluster GC6. Cluster GC5, while not very representative, is mainly shared by two different authors that use the same choices.

A closer look at the tables of the centroids (Tables 5 and 7) and the tables of the numbers of consistent individuals (Tables 9 and 10) reveals that all the clusters characterized by the use of the definite determiner (GC2 and GC5) are the ones with a greater percentage of referring expressions written by consistent individuals. This could be due to the appropriateness of definite determiners to lexicalize all kinds of referring expressions. For example, expressions where comparatives or superlatives must to be lexicalized (e.g., *the smaller of the two blue fans*). However, the use of indefinite determiners in these and other similar situations do not make it possible to generate correct or natural expressions. Thus, it is more difficult to stay true

Table 11. Use of position, orientation, and non-positional attributes in the attribute set.

	position	orientation	non-positional attributes	none
Furniture	150 (38%)	127 (32%)	352 (88%)	5 (1%)
People	132 (39%)	9 (3%)	251 (74%)	0

to choices that include such constraints. A similar explanation can be given to the absence of consistent individuals in cluster GC4. The use of non-structured expressions may be not suitable to lexicalize all the information needed to refer to an object under any circumstances. As a result, nobody stayed true to these choices through all their referring expressions. Cluster GC6 only contained two expressions, not enough for having a consistent author.

The results reveal that variability is present in the lexicalization of referring expressions, even in a very restricted context such as the one considered here and with this few annotation elements. This variability gives rise to interesting manifestations, such as popular and rare preferences, or consistent and non-consistent individuals. In the next subsection, we will go deeper into the description of the variability of the lexicalization in the TUNA corpus focusing on more annotation elements.

5.3 Finding Groupings According to Specific Grammatical Choices

The previous analysis showed that it is possible to find very homogeneous groupings of general grammatical choices for lexicalized referring expressions. However, the expressions in each cluster showed a degree of heterogeneity in terms of other grammatical choices that were not taken into account in the analysis carried out in Section 5.2. These grammatical choices are specific to the attributes chosen to refer to the intended referent in each expression. The main attributes we considered are position and orientation and the rest of attributes (henceforth referred to as 'non-positional' attributes for the sake of clarity). This division is due to the high variability found in the lexicalization of attributes that express position and orientation (see Table 2 in Section 1) explained in Section 4.1. Obviously, not all the expressions in the corpus use the same attributes to refer to the intended referent and some of them use more than one attribute. Table 11 shows the number of expressions in each domain that use each of these three kinds of attributes. As can be seen, most of the expressions used non-positional attributes. Around 40% of expressions in both domains used the position in the grid of the experiment to refer to the person or piece of furniture. Interestingly, orientation was seldom used in the people domain, while it was more popular in the furniture domain.

In this section we will aim to find groupings of expressions that share similar grammatical choices to represent one of the attributes mentioned. We will show

Table 12. Frequency of use of the different ways of expressing non-positional attributes in the people domain.

Premodifier	With	In	Relative Cl.	Gerund Ph.	Indicative Cl.
62 (25%)	199 (79%)	5 (2%)	4 (2%)	26 (10%)	1 (<1%)

the results obtained for the cluster analysis of the grammatical choices when using non-positional attributes in the people domain and when using position in the furniture domain. The other possible analyses are position and orientation in the people domain, and orientation and non-positional attributes in the furniture domain. Orientation cannot be analyzed in the people domain because only 9 referring expressions use orientation. Interestingly, the use of non-positional attributes in the furniture corpus shows no variability, because 99% of the expressions represent the attribute using an adjective as premodifier. Due to this lack of variability, it made no sense to carry out further cluster analysis for this case. The conclusion that can be drawn is that the expression of non-positional attributes for "things" does not exhibit the same level of richness as in the case of people. The other two clustering analyses that were carried out were the use of orientation in the furniture domain and the use of position in the people domain. These analyses provide descriptive information about the referring expressions in the TUNA corpus, but do not provide further insight about lexical choices beyond that provided by the two analyses that will be detailed in the next sections.

5.3.1 Non-positional Attributes in the People Domain

In the people domain, 251 out of 340 referring expressions use attributes different from orientation and position to refer to the required person. The non-positional attributes used are mainly age, hair, hair colour, beard, glasses, and clothes. The annotation elements that represent the grammatical choices for expressing a non-positional attribute are the ones enumerated in Section 4.1.4.

Table 12 shows the frequency of use for each of these grammatical choices for expressing a non-positional attribute in this domain. The most popular way of expressing attributes is the use of "with" (79%), while the premodifiers and the use of gerunds are much less frequent (25% and 10%, respectively). The other ways of representing non-positional attributes are barely used.

Table 13 shows the distribution of the 251 expressions that use non-positional attributes among the six clusters identified in Section 5.2. The percentage that these referring expressions represent in the original cluster is shown in parentheses. The table shows that in all the clusters a fair share of expressions use non-positional attributes to refer to a person. In cluster GC4, which is the cluster with no structured expressions, the percentage is smaller (55%) because this cluster contains

Table 13. Distribution of the expressions with non-positional attributes in the six clusters found in the people domain using the variables of the general grammatical choices.

GC1	GC2	GC3	GC4	GC5	GC6
96 (73%)	88 (70%)	43 (98%)	12 (55%)	10 (71%)	2 (100%)

Table 14. Centroids of each cluster of the people domain.

	AC1	AC2	AC3	AC4	AC5	AC6	AC7
DET-DEF	0 (70%)	0 (67%)	**1** (100%)	0 (86%)	0 (100%)	0 (100%)	0 (100%)
DET-INDEF	0 (97%)	0 (93%)	0 (100%)	0 (68%)	**1** (100%)	**1** (100%)	0 (100%)
STRUCT-PRED	0 (97%)	0 (100%)	0 (89%)	0 (100%)	0 (75%)	0 (97%)	0 (100%)
STRUCT-NOM	**1** (73%)	**1** (100%)	**1** (89%)	**1** (100%)	**1** (100%)	**1** (100%)	**1** (95%)
E-O-A-PREMOD	**1** (97%)	0 (87%)	0 (100%)	**1** (100%)	**1** (75%)	0 (100%)	0 (100%)
E-O-A-WITH	0 (100%)	0 (100%)	**1** (99%)	**1** (100%)	0 (100%)	**1** (100%)	**1** (97%)
E-O-A-IN	0 (97%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (94%)
E-O-A-REL	0 (100%)	0 (100%)	0 (100%)	0 (100%)	**1** (100%)	0 (100%)	0 (100%)
E-O-A-GER	0 (100%)	**1** (100%)	0 (97%)	0 (89%)	0 (100%)	0 (100%)	0 (90%)
E-O-A-INDIC	0 (100%)	0 (100%)	0 (99%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)

many referring expressions such as *top left* or *second top* where only the position is used.

The k-means clustering method was applied to find groupings of expressions with similar grammatical choices. The variables considered were the four annotation elements that represent grammatical choices and the six that represent different ways of expressing the non-positional attributes of an element. After different attempts, the number of clusters considered was $k = 7$ because it produced meaningful and homogeneous clusters.

Table 14 shows the centroid values for the seven resulting clusters (the percentage of observations of the cluster whose value is equal to that of the centroid is shown in parentheses). The number of referring expressions in each cluster and the variability measures are shown in Table 15.

The sets of similar choices found in this new clustering for the people domain are:

- **Cluster AC1** mainly has nominal structure, uses no determiners and represents the non-positional attributes as premodifiers, like in *balding grey haired grey bearded man* or *white bearded man*. The cluster has 30 expressions. However, the cluster is heterogeneous. 27% of the expressions have no structure

Table 15. Number of observations, within-cluster variability and mean within-cluster variability in each cluster of the people domain.

	AC1	AC2	AC3	AC4	AC5	AC6	AC7
# obs.	30 (12%)	15 (6%)	80 (32%)	28 (11%)	4 (1%)	32 (13%)	62 (25%)
total var.	21	8	24	16	2	1	15
mean var.	0.7	0.53	0.3	0.57	0.5	0.03	0.23

at all (e.g., *glasses and white beard*) and 30% of the expressions use a definite determiner (*the white-bearded man on the left*).

- **Cluster AC2** has only 15 expressions. They have nominal structure and represent the non-positional attribute by using a gerund form. The use of determiners in the cluster is heterogeneous. 13% of the referring expressions have an non-positional attribute represented as a premodifier. Examples of this cluster include: *man wearing glasses* or *elderly gentleman wearing heavy framed spectacles*. The gerund is in all the cases "wearing" and always refer to glasses.
- **Cluster AC3** corresponds to expressions with nominal structure, which use the definite determiner and represent non-positional attributes using "with". This is the biggest cluster with 80 expressions. Examples of this cluster are *the man with a beard* and *the man with black hair and glasses*. 11% of the referring expressions in this cluster have predicative instead of nominal structure, for example, *the guy with glasses in the middle of the top row is in red box*.
- **Cluster AC4** has 28 expressions with nominal structure that represent at least two non-positional attributes, one of them using "with", while the other is represented as a premodifier. Some examples of this cluster are *dark haired man with grey beard* or *a young man with glasses*. 11% of the expressions also represent a third non-positional attribute by using a gerund form as in *elderly gentleman with a white beard wearing spectacles*. The use of the determiner is not homogeneous in the cluster.
- **Cluster AC5** has only four referring expressions, all of them written by the same author. The referring expressions mainly have nominal structure, use the indefinite determiner and express one non-positional attribute by using a relative clause and in 75% of the cases a second non-positional attribute as a premodifier. An example is *a young man who is wearing glasses and who is looking forward*. This cluster represents the set of choices of one person that exhibits a great deal of consistency. However, these choices are shared by no one else in the corpus, which suggests that they are not usual.
- **Cluster AC6** represents the expressions with nominal structure and an indefinite determiner that use "with" for the non-positional attribute. This cluster has 32 expressions such as *a man with glasses and black hair* or *a man with glasses and a white beard*.

- **Cluster AC7** is characterized by nominal structure, the use of no determiners and the use of "with" to represent the non-positional attributes. It is the second biggest cluster with 62 referring expressions. Some expressions found in this cluster are: *man with grey beard* or *male subject on bottom with beards*. Some of the expressions of the cluster use an additional non-positional attribute using a gerund form (*man with beard and wearing glasses*) or with "in" (*man in tie and glasses with dark jacket*).

The consistency of the clusters does not offer as much information as in the case of general grammatical choices. This is due to the fact that the number of referring expressions written by each person using non-positional attributes is not always the same. In addition, although a non-positional attribute can be expressed in different ways, they sometimes depend on the attribute itself; for example, there is no direct way to say "with glasses". This makes more difficult to find consistency. As a result, if we study the data looking for consistent individuals that have written more than one referring expression with non-positional attributes, we find that there are only 10 consistent individuals. Interestingly, eight of those individuals belong to cluster AC3 (for a total of 34 referring expressions out of 80 referring expressions in that cluster) and two individuals belong to cluster AC7 (for a total of 10 referring expressions out of 62). It is not surprising to find consistent individuals in the two biggest clusters. Most of the remaining individuals wrote referring expressions that belong to two clusters. Interestingly, the individual that wrote the four referring expressions that make up cluster AC5 is not a consistent individual, because he wrote two other referring expressions that belong to other clusters.

5.3.2 Position in the Furniture Domain

In the furniture domain, 150 expressions out of 398 (i.e., close to 40% of the corpus) include the position of the referred piece of furniture in the experiment grid. They reveal that there is a great variety of ways to express the position. The new annotations that represent the grammatical choices for expressing the position are the ones enumerated in Section 4.1.2.

First we studied the frequency of use for each of the position variables employed. Table 16 shows their absolute and relative frequency. It is interesting to see that close to 90% of the expressions represented the position using top-bottom (left-right) and 33% of the expressions used rows and columns to refer to the piece of furniture indicated. On the contrary, the use of cardinal points and cardinal numbers in the corpus is negligible as they were used only once.

Table 17 shows the distribution of the 150 expressions that use position among the six clusters identified in Section 5.2. The percentage that these referring expressions represent in the original clusters is shown in parentheses. In the two biggest clusters, GC1 and GC2, 35% and 44% of the expressions use position. However, only 4% of the expressions of cluster GC3 (that is, 2 out of 53) use position. This may be due to the fact that cluster GC3 is characterized by the use of indefinite determiners and this kind of determiners is not suitable to express position. On the contrary,

Table 16. Frequency of use of the different ways of representing position in the furniture domain.

RowColumn	TopBottom	Cardinal	Card-Num	Ord-Num	Relation
50 (33%)	131 (87%)	1 (< 1%)	1 (<1%)	17 (11%)	10 (7%)

Table 17. Distribution of the expressions using position in the six clusters found in the furniture domain using the variables of general grammatical choices.

GC1	GC2	GC3	GC4	GC5	GC6
52 (35%)	65 (44%)	2 (4%)	17 (55%)	13 (93%)	1 (33%)

93% of the expressions of cluster GC5 use position and cluster GC5 is characterized by the use of the definite determiner. In cluster GC4, the cluster without structure, we find that 55% of expressions use position. Only 1 out of 3 expressions of cluster GC6 used position, but given such a small number, the result is not significant.

We applied a k-means clustering method to group referring expressions with similar grammatical and position choices. The variables considered included: the four variables that represent grammatical choices, and the six variables that represent different ways of expressing the position of an element in the grid of the experiment. The clustering was carried out for different values of k, i.e. number of clusters. After a careful analysis, we considered that the most meaningful results were obtained for $k = 6$. Table 18 shows the centroid values for the six clusters obtained (the percentage of observations of the cluster whose value is equal to that of the centroid is shown in parentheses). The number of referring expressions in each cluster and the variability measures are shown in Table 19.

The sets of similar choices found in this new clustering for the furniture domain are:

- **Cluster PC1** is a small cluster that represents the expressions with nominal structure that express the position using a relation (e.g., *the red chair above the blue one*). Some of these expressions use definite or indefinite determiners and some of them use no determiner at all. More interestingly, 56% of the expressions represent the position using top-bottom (e.g., *red sofa bottom right corner of screen next to blue sofa and below green desk*).
- **Cluster PC2** represents a quite homogeneous way of referring to objects that consists of nominal structure, no determiners, and position represented with top-bottom. It is the second biggest cluster with 38 referring expressions. Prototypical examples of PC2 are *blue chair on top* or *top blue fan*. However, in this cluster 24% of the expressions also use a reference to the row or the column, such as in the case of *green desk in middle row*.

Table 18. Centroids of each cluster of the furniture domain.

	PC1	PC2	PC3	PC4	PC5	PC6
DET-DEF	0 (78%)	0 (100%)	**1** (100%)	0 (59%)	0 (82%)	**1** (100%)
DET-INDEF	0 (88%)	0 (100%)	0 (100%)	0 (95%)	0 (100%)	0 (100%)
STRUCT-PRED	0 (100%)	0 (100%)	**1** (100%)	0 (55%)	0 (100%)	0 (100%)
STRUCT-NOM	**1** (100%)	**1** (100%)	0 (100%)	0 (100%)	**1** (71%)	**1** (100%)
POS-ROWCOL	0 (89%)	0 (76%)	**1** (100%)	0 (95%)	**1** (100%)	0 (70%)
POS-TBLR	**1** (56%)	**1** (100%)	**1** (75%)	**1** (100%)	0 (82%)	**1** (100%)
POS-CARD-POINT	0 (100%)	0 (100%)	0 (75%)	0 (100%)	0 (100%)	0 (100%)
POS-CARD-NUM	0 (100%)	0 (97%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)
POS-ORD-NUM	0 (100%)	0 (100%)	0 (100%)	0 (100%)	**1** (100%)	0 (100%)
POS-WITH-REL	**1** (100%)	0 (100%)	0 (100%)	0 (100%)	0 (100%)	0 (98%)

Table 19. Number of observations, within-cluster variability and mean within-cluster variability in each cluster of the furniture domain.

	PC1	PC2	PC3	PC4	PC5	PC6
# obs.	9 (6%)	38 (25%)	4 (3%)	22 (15%)	17 (11%)	60 (40%)
total var.	8	10	2	21	11	19
mean var.	0.89	0.26	0.5	0.95	0.65	0.32

- **Cluster PC3** is very small (only 4 expressions), but all of them are homogeneous in terms of general grammatical choices (they use predicative structure and the definite determiner) and all of them use row-column to represent the position (e.g., *the chair in the top row is in a red box*). However, this cluster does not represent shared choices as all its expressions were written by the same individual.
- **Cluster PC4** has as prototype a referring expression with no structure and no determiners and that uses top-bottom to represent the position. For example, expressions such as *right bottom* or *blue chair center*. However, expressions with the definite determiner and predicative structure are also an important part of the cluster (e.g., *the chair at the top right is surrounded by a border* or *the red sofa at the top left is surrounded by a border*). These two different kinds of expressions represent the bulk of the 22 referring expressions in the cluster.
- **Cluster PC5** mostly represents expressions with nominal structure that do not use determiners and where the position is represented with row-column and ordinal numbers (e.g., *first picture on third row*). However, 29% of the expressions have no structure at all (e.g *grey desk first row*). Note that the

grammatical difference between the two kinds of expressions is the use of a preposition. 18% of the referring expressions in the cluster use the definite determiner (e.g., *the grey chair in the second row*). Regarding position, 18% of the expressions in the cluster use top-bottom, in addition to the features mentioned above, to represent it (e.g., *the red chair in the middle of the first row* or *top row first fan only*).

- **Cluster PC6** uses definite determiners and nominal structure and expresses the position using top-bottom. It is the biggest cluster obtained, with 60 expressions. An example of referring expression included in this cluster is *the one in the middle top*. 30% of the expressions in this cluster also use row and column (e.g., *the blue chair in the bottom row*). This cluster represents almost all expressions in cluster GC2 (60 out of 65) obtained by considering only the grammatical choices (see Table 16). Thus, it can be said that the joint use of *definite determiner*, *nominal structure* and *position using top-bottom* is very popular in the corpus.

As happened in Section 5.3.1, the study of the consistency of these clusters offers less information than in the general grammatical choices analysis. We found 11 individuals who wrote all their referring expressions with position using the same set of choices. Six of these individuals belong to cluster PC6, which is the biggest cluster. Those individuals wrote 32 of the referring expressions of the cluster, which means more than 50% of the cluster referring expressions. The rest of the consistent individuals belong to middle size clusters: two to cluster PC2 (for a total of 7 referring expresions), one to cluster PC4 (5 referring expressions) and one to cluster PC5 (3 referring expressions). An individual can be consistent with 3 or 5 referring expressions if they are the only ones in which he used the position attribute we are studying. As happened in the study of attributes for the people domain, the individual who wrote all the referring expressions of a small cluster, PC3 in this case, is not a consistent individual, because he wrote other expressions that belong to other clusters. In this case, it is also observed that most individuals write referring expressions switching between two different sets of lexical choices.

6 Discussion

The work presented in this paper showed that we can find sets of homogeneous lexical choices in the TUNA corpus. These sets are mainly shared by different authors, and we have also shown that some people switch between different sets. Additionally, we have seen that these lexical choices are not equally popular, but some of them are more used than others.

The annotation process we carried out added new lexical information to the corpus. This expanded annotation of TUNA is expected to provide a useful resource for further research on lexicalization of referring expressions. Many previous research efforts have relied on the original TUNA corpus for the choice of conceptual information when generating referring expressions. These efforts range from several REG algorithms implemented based on the findings of the corpus, including the participants in the REG Shared Tasks (Gatt 2007; Belz and Gatt 2007;

Gatt *et al.* 2008; Gatt *et al.* 2009), to theoretical works assessing the psychological basis of existing algorithms (Belz and Gatt 2008; van Deemter *et al.* 2012). Thanks to our work it will be possible to extend these efforts in the choice of lexical forms for the conceptual information chosen. The extended corpus is available in `http://nil.fdi.ucm.es/index.php?q=node/567`.

We have also exploited the corpus annotation with data analysis methods. The aim was to study the lexical variability of referring expressions in the TUNA corpus. The results obtained in the cluster analysis show that variability is a rich phenomenon with interesting manifestations that could be taken into account in NLG. An interesting finding was that not only does each person present different choices in different expressions, but that several sets of these choices are shared by different people. In addition, although many people are mostly true to the use of some set of lexical choices, others switch between different sets. This evidence, although intuitively plausible, had not been empirically studied. The explanation is probably rooted in the field of Psycholinguistics and goes beyond the scope of the present paper.

The obtained clusters represent sets of choices in terms of grammatical structure, use of determiners, choice of expressions for certain attributes, and so on. Any REG algorithm could benefit from this data to add variability to its responses. Depending on the paradigm used in the implementation of the algorithm, it would be possible to create rules that represent the choices, or to train the system using specific clusters. The Case-Based Reasoning (CBR) paradigm (Aamodt and Plaza 1994) is specially suitable in this case because it is possible to generate expressions that take into account already known lexical choices by narrowing the case base to only those referring expressions that use the same lexical features that we want to reproduce.

An interesting question is whether sets of lexical choices can be exploited to improve the performance of REG algorithms. To answer this question we have performed a preliminary experiment of referring expression generation considering the information about lexical choices obtained using clustering. For that purpose we have used a previously implemented REG system (Hervás 2009) that relies on CBR. The results show that the performance of the system is greatly improved when considering the lexical preferences from the clusters, obtaining referring expressions that match the variability of the human-generated expressions of the TUNA corpus.

In our CBR-based REG system a case is composed of two parts: a problem description and a solution. The problem description is a referring expression plan which lists the set of attributes that need to be mentioned. The goal of the CBR system is to determine how these particular attributes will be lexicalized. Therefore, the final lexicalization is the case solution. Thus, the case base is a set of referring expressions, each of them containing a plan (attribute set) and a solution (final lexicalization). The system works as follows:

1. When a new referring expression needs to be lexicalized, the system is presented with the attribute set for the referent.
2. Then, the system retrieves the expression in the case base whose attribute set is most similar to the new one.

Table 20. Comparison of string-edit distance results obtained using the original corpus vs. the clusters that define general choices and specific choices. Differences to the original corpus are significant at p<0.001 (*).

	Original Corpus	**General Choices**	**Specific Choices**
Furniture	5.74	4.16 *	–
	6.59	–	5.34 *
	–	4.85	5.13
People	6.51	4.85 *	–
	6.23	–	4.23 *
	–	4.61	4.02 *

3. Finally, the expression retrieved is used to lexicalize the new referring expression by reusing its syntactic and lexical forms.

Our experiment assesses if using lexical choices as an additional input improves the results obtained by the system. In order to do so, we will compare the results obtained by the CBR system using case bases created in two different ways: by taking into account the clusters (using referring expressions that are written with the same set of lexical choices as case base) and by not considering them (using the whole corpus as case base). In addition, we will compare the performance of the system using the clusters of general grammatical choices described in Section 5.2, and afterwards using the clusters of specific grammatical choices shown in Section 5.3

For both approaches we performed a k-fold cross validation with k=10. The results obtained were measured using string-edit distance (Levenshtein 1966). The string-edit distance between two strings is given by the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single word. This measure is typically used in the literature for assessing the similarity between original and generated referring expressions, and it requires a gold standard for comparison with the referring expression that is being evaluated. We have considered as the gold standard the lexicalization that appears in the TUNA corpus for the same expression that is being lexicalized. The Wilcoxon Signed Ranks Test was performed in order to test whether differences between results were statistically significant. Two other metrics were also used to measure the performance of the system: BLEU (Papineni *et al.* 2002) and ROUGE (Lin and Och 2004), but the results were similar to the ones obtained using string-edit distance, and for the sake of brevity have not been included in the paper.

The results for the general and specific grammatical choices are shown in Table 20. Three different comparisons are shown for each domain: original corpus vs.

general choices, original corpus vs. specific choices, and general vs. specific choices. The results of the system using the original corpus are rather poor when it comes to matching the specific expressions used in the corpus. We consider this to be due to the nature of the corpus, which exhibits a broad variety of expression. In this way, the same set of attributes can be lexicalized in a varied set of forms, and most times the chosen lexicalization is correct but does not correspond to the original one in the corpus. However, the results are improved when the referring expressions are generated after considering lexical choices. In those situations, we are narrowing some of the syntactic and lexical decisions as we are considering some of the stylistic preferences used when the expression was originally generated. Specific choices improve the results obtained with general choices in the people domain, but not in furniture. This result was not completely surprising because the clusters obtained in this case were quite heterogeneous, with numerous referring expressions in the groups that were not similar to the centroid.

7 Conclusions and Future Work

This paper studies the influence of lexical variation in a specific task of the NLG process: the lexicalization of referring expressions. Data presented here displays the high lexical variability of natural language in general and the TUNA corpus in particular. Our results show that personal choices may be adequately captured in terms of syntactic and lexical features. Clustering techniques may be used to identify consistent subsets of choices, and these sets may be applied to the functioning of REG algorithms. By means of this two-step approach (annotation and analysis) we have confirmed that lexical variability across a collection of examples is not only due to the fact that different examples are created by different people who make different choices, because groups of people may also share similar choices.

In summary, the conclusions we can draw from this paper are:

- We can find sets of homogeneous lexical choices in the TUNA corpus.
- These sets are mainly shared by different authors.
- Not all the authors are loyal to their personal choices, some people switch between different sets of choices.
- All lexical choices are not equally popular, some of them are more used than others.

The work presented in this paper opens several lines of future work. For example, the annotation process may be partially automated using syntactic analysis techniques. Some features like the presence of determiners or the grammatical structure of the sentence would be easily annotated, although more complex ones may be more difficult to detect. In any case, the partial or full automation of the annotation process would improve the detection of sets of similar lexical choices even if manual validation is required later. In addition, if more information about the speaker were available, it would be interesting to study whether there are any correlations between the user profile and his choices when generating language.

An interesting experiment would be to check whether an author that makes

certain lexical choices also prefers to read referring expressions created using the same choices. As we have not access to the authors of the original TUNA corpus this experiment cannot be carried out, but in future work it would be possible to create a new small corpus of referring expressions, apply our methodology in three steps, and confront the authors with referring expressions created with the same choices they have used. This would be a possible application of the ideas of this paper in Accommodation Theory (Giles *et al.* 1991).

From the point of view of the clustering analysis, the obtained clusters represent lexical choices shared by a set of authors for lexicalizing similar conceptual information. However, the clusters are also influenced by some attributes that must usually be expressed in specific ways. In future work it would be interesting to study the influence of correct and incorrect lexicalizations of different attributes on the obtained clusters.

Finally, we firmly believe that the approach presented in this paper transcends the limited example laid out here and can be helpful to other natural language processing approaches. The general procedure applied in this paper may generalize to a broader range of tasks as well as referring expression generation. Overall, this procedure follows the trend of learning decisions from sets of data rather than handcrafting solutions based on prior generic knowledge. The generic procedure to be applied for a specific task X could be summarized as follows: annotate a corpus with information about specific choices made during the performance of task X, analyze this annotated corpus to study how observed choices can be grouped into sets of choices for generation, and assess the influence of considering such sets in the performance of an existing system that undertakes task X as part of its operation.

Given that the final aim of NLG is to produce human-like language and that personal choices are clearly present in language, acknowledging their importance and trying to incorporate them into NLG systems is a path that should be followed.

Acknowledgements

This work is partially funded by ConCreTe. The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

References

Aamodt, A. and Plaza, E. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications 7*: 39-59.

Artstein, R. and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics 34*: 555-596.

Belz, A. and Gatt, A. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *2nd UCNLG Workshop: Language Generation and Machine Translation*, pp. 75-83, Copenhaguen, Denmark.

Belz, A. and Gatt, A. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Columbus, Ohio. Association for Computational Linguistics.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Bohnet, B. 2008. The fingerprint of human referring expressions and their surface realization with graph transducers. In *Referring Expression Generation Challenge 2008, 5th International Natural Language Generation Conference*, pp. 207-210, Salt Fork, Ohio. Association for Computational Linguistics.

Bohnet, B. 2009. Generation of referring expression with an individual imprint. In *Generation Challenges 2009, European Natural Language Generation Conference*, pp. 185-186, Athens, Greece. Association for Computational Linguistics.

Brennan, R. L. and Prediger, D. J. 1981. Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement 41:* 687-699.

Dale, R. and Viethen, J. 2009. Referring expression generation through attribute-based heuristics. In *12th European Natural Language Generation Conference*, pp. 85-65, Athens, Greece. Association for Computational Linguistics.

Dale, R. and Viethen, J. 2010. Empirical Methods in Natural Language Generation. Chapter Attribute-Centric Referring Expression Generation, pp. 163-179. Berlin, Heidelberg: Springer-Verlag.

di Fabbrizio, G., Stent, A. and Bangalore, S. 2008. Referring expression generation using speaker-based attribute selection and trainable realization. In *Referring Expression Generation Challenge 2008, 5th International Natural Language Generation Conference*, pp. 211-214, Salt Fork, Ohio. Association for Computational Linguistics.

Gatt, A. 2007. *Generating Coherent References to Multiple Entities.* Ph.D. thesis, University of Aberdeen, UK.

Gatt, A., van der Sluis, I. and van Deemter, K. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *11th European Workshop on Natural Language Generation*, pp. 49-56, Germany. Association for Computational Linguistics.

Gatt, A., van der Sluis, I. and van Deemter, K. 2008. *XML format guidelines for the TUNA corpus*. Technical Report, University of Aberdeen.

Gatt, A., Belz, A. and Kow, E. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *5th International Conference on Natural Language Generation*, pp. 198-206, Ohio, USA. Association for Computational Linguistics.

Gatt, A., Belz, A. and Kow, E. 2009. The TUNA-REG Challenge 2009: Overview and evaluation results. In *12th European Workshop on Natural Language Generation*, pp.

174-182, Athens, Greece. Association for Computational Linguistics.

Giles, H., Coupland, J. and Coupland, N. 1991. *Contexts of Accommodation: Developments in Applied Sociolinguistics*. New York: Cambridge University Press.

Hervás, R. 2009. *Referring Expressions and Rhetorical Figures for Entity Distinction and Description in Automatically Generated Discourses*. Ph.D. thesis, Universidad Complutense de Madrid, Spain.

Hervás, R., Francisco, V. and Gervás, P. 2013. Assessing the influence of personal preferences on the choice of vocabulary for natural language generation. *Information Processing and Management*, *49*: 817-832.

Jain, A. K., Murty, M. N. and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys*, *31*: 264-323.

Krahmer, E. and van Deemter, K. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*: 173-218.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*: 707-710.

Lin, C. and Och, F. J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain. Association for Computational Linguistics.

MacQueen, J. B. 1967. Some methods for classification and analysis of multiVariate observations. In L. M. L. Cam and J. Neyman (Eds.), *5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*1 281-297. University of California Press

Mairesse, F. and Walker, M. A. 2011. Controlling user perceptions of linguistic style: trainable generation of personality traits. *Computational Linguistics*, *37*: 455-488.

Paiva, D. and Evans, R. 2005. Empirically-based control of natural language generation. In *43rd Annual Meeting on Association for Computational Linguistics*, pp. 58-65, Ann Arbor, Michigan. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. 2002. BLEU: A method for automatic evaluation of machine translation. In *40th Annual Meeting on Association for Computational Linguistics*, pp. 311-318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Power, R., Scott, D. and Bouayad-Agha, N. 2003. Generating texts with style. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 93-105. Berlin, Heidelberg: Springer Verlag.

Randolph, J. J. 2005. Free-marginal multirater Kappa: An alternative to fleiss' fixed-marginal multirater Kappa. In *Joensuu University Learning and Instruction Symposium*.

Randolph, J. J. 2008. Online Kappa Calculator. `http://aiweb.techfak.uni-bielefeld.de/content/bworld-robot-control-software/`

Reiter, E. and Dale, R. 2000. *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Reiter, E., Sripada, S., Hunter, J., Yu, J. and Davy, I. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, *167*: 137-169.

Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, *19*: 321-325.

van Deemter, K., van der Sluis, I. and Gatt, A. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *4th International Conference on Natural Language Generation (Special Session on Data Sharing and Evaluation)*, pp. 130-132, Sydney, Australia.

van Deemter, K., Gatt, A., van der Sluis, I. and Power, R. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, *36(5)*: 799-836.

Viethen, J. and Dale, R. 2010. Speaker-dependent variation in content selection for referring expression generation. In *8th Australasian Language Technology Workshop*, pp. 81-89.