

TUNA Corpus *Extended* - Annotation Guide

Raquel Hervás, Javier Arroyo, Virginia Francisco, Federico Peinado, Pablo Gervás

Departamento de Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid, 28040, Madrid, Spain

The TUNA Reference Corpus [1,2] is a semantically and pragmatically transparent corpus of identifying references to objects in visual domains. It was constructed via an on-line experiment, and has since been used in a number of evaluation studies on Referring Expressions Generation. More information about the TUNA corpus can be found in <http://inf.abdn.ac.uk/research/tuna/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. Therefore, the TUNA Corpus *Extended* is an extended version of the original corpus in which some annotations related to specific aspects of referring expressions are added.

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 a high variation and a significant number of mentions in the original corpus. To facilitate later data analysis, all the annotations are represented by binary values; that is, **yes** or **no** are the only possible values.

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. Depending on the type of determiner, there are two different elements in the annotation:
 - **DETERMINER-DEF** (e.g., *the man*).
 - **DETERMINER-INDEF** (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:
 - **STRUCT-PRED.** The referring expression has a verbal structure, that is, it is a predicative phrase (e.g., *the chair is gray*).
 - **STRUCT-NOM.** 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.

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.

- **POS-CARDINAL-NUMBERS.** The position of the referent is described using cardinal numbers (e.g., *the desk in row two*).
- **POS-CARDINAL-POINT.** The position of the referent is described using cardinal points (e.g., *man in the north part of the screen*).
- **POS-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.
- **POS-WITH-RELATION.** 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*).
- **POS-ROWCOLUMN.** 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*).
- **POS-TOPBOTTOMLEFTRIGHT.** 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*).

If the referring expression presents two or more of the defined situations simultaneously each of the previous elements is annotated accordingly.

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*”).

- **EXPRESSION-OF-ORIENTATION-DIRECT.** The orientation of the referent is described directly, without using any kind of connectives or auxiliary phrases (e.g., *the backview desk*).
- **EXPRESSION-OF-ORIENTATION-GERUND.** The orientation of the referent is described using a gerund phrase (e.g., *the chair facing away from me*).
- **EXPRESSION-OF-ORIENTATION-PARTICIPLE.** The orientation of the referent is described using a participle phrase (e.g., *the chair shown from its side*).
- **EXPRESSION-OF-ORIENTATION-RELATIVE.** 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*).
- **EXPRESSION-OF-ORIENTATION-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*).

If the referring expression presents two or more of the defined situations simultaneously each of the previous elements is annotated accordingly.

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.

- **EXPRESSION-OF-ATTRIBUTE-PREMOD.** The attributes of the referent are described using a premodifier, without using any kind of connectives or auxiliary phrases (e.g., *the bald man*).
- **EXPRESSION-OF-ATTRIBUTE-GERUND.** The attributes of the referent are described using a gerund phrase (e.g., *the man wearing glasses*).
- **EXPRESSION-OF-ATTRIBUTE-INDICATIVE.** The attributes of the referent are described using an indicative clause (e.g., *it is a fan in the middle and it is small*).
- **EXPRESSION-OF-ATTRIBUTE-RELATIVE.** The attributes of the referent are described using a relative clause (e.g., *the man who has black hair*).
- **EXPRESSION-OF-ATTRIBUTE-IN.** The attributes of the referent are connected to the referent by the preposition ‘in’ (e.g., *The man in suit*).
- **EXPRESSION-OF-ATTRIBUTE-WITH.** The attributes of the referent are connected to the referent by the prepositions ‘with’ or ‘without’ (e.g., *the man with glasses*).

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.

Figure 1 presents an example of a referring expression from the original TUNA corpus with the new annotations.

```

<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. 1. Example of referring expression from the TUNA corpus re-annotated with lexical information.

2 Notes on the Annotation Process

Sometimes the referent does not appear in the sentence we are annotating. Usually in those cases the sentence is considered as not having a valid syntactical structure (e.g. “right middle”).

Sentences as “beard, glasses, old” represent an interesting annotation issue: they have been annotated as direct expression of attributes of the referent, but

some of these words (“beards” and “glasses”) are, in fact, parts of the referent. Our decision is that everything that appears in the attribute set of the original corpus cannot be considered as a part but as an attribute of the referent.

Ambiguity is inherent to natural language, and the annotators may find tricky situations during the annotation process. For example, “. . . with the seat on the right”, where “on the right” can be considered connected with the seat or with the referent itself. Here the advice is to annotate what seems more plausible. Sometimes the word “frontal” presents a similar problem, because it can be considered an expression of orientation or an expression of position (although obviously the more plausible is to consider it an expression of orientation).

References

1. Van Deemter, K., Van Der Sluis, I., Gatt, A.: Building a Semantically Transparent Corpus for the Generation of Referring Expressions. In: Proceedings of the 4th International Conference on Natural Language Generation (Special Session on Data Sharing and Evaluation), INLG-06. (2006)
2. Gatt, A., Van Der Sluis, I., Van Deemter, K.: Evaluating Algorithms for the Generation of Referring Expressions Using a Balanced Corpus. In: Proc. of the 11th European Workshop on Natural Language Generation (ENLG 07). (2007)