The Prevalence of Descriptive Referring Expressions in News and Narrative

Raquel Hervás
Departamento de Ingenieria
del Software e Inteligencía Artificial
Universidad Complutense de Madrid
Madrid, 28040 Spain
raquelhb@fdi.ucm.es

Mark Alan Finlayson
Computer Science and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, 02139 USA
markaf@mit.edu

Abstract

Generating referring expressions is a key step in Natural Language Generation. Researchers have focused almost exclusively on generating *distinctive* referring expressions, that is, referring expressions that uniquely identify their intended referent. While undoubtedly one of their most important functions, referring expressions can be more than distinctive. In particular, *descriptive* referring expressions – those that provide additional information not required for distinction – are critical to fluent, efficient, well-written text. We present a corpus analysis in which approximately one-fifth of 7,207 referring expressions in 24,422 words of news and narrative are descriptive. These data show that if we are ever to fully master natural language generation, especially for the genres of news and narrative, researchers will need to devote more attention to understanding how to generate descriptive, and not just distinctive, referring expressions.

1 A Distinctive Focus

Generating referring expressions is a key step in Natural Language Generation (NLG). From early treatments in seminal papers by Appelt (1985) and Reiter and Dale (1992) to the recent set of Referring Expression Generation (REG) Challenges (Gatt et al., 2009) through different corpora available for the community (Eugenio et al., 1998; van Deemter et al., 2006; Viethen and Dale, 2008), generating referring expressions has become one of the most studied areas of NLG.

Researchers studying this area have, almost without exception, focused exclusively on how to generate *distinctive* referring expressions, that is, referring expressions that unambiguously identify their intended referent. Referring expressions, however, may be more than distinctive. It is widely acknowledged that they can be used to achieve multiple goals, above and beyond distinction. Here we focus on *descriptive* referring expressions, that is, referring expressions that are not only distinctive, but provide additional information not required for identifying their intended referent. Consider the following text, in which some of the referring expressions have been underlined:

> *Once upon a time there was a man, who had three daughters. They lived in a house and their dresses were made of fabric.*

While a bit strange, the text is perfectly well-formed. All the referring expressions are distinctive, in that we can properly identify the referents of each expression. But the real text, the opening lines to the folktale *The Beauty and the Beast*, is actually much more lyrical:

> *Once upon a time there was a* ***rich merchant****, who had three daughters. They lived in a* ***very fine*** *house and their* ***gowns*** *were made of* ***the richest*** *fabric* ***sewn with jewels****.*

All the boldfaced portions – namely, the choice of head nouns, the addition of adjectives, the use of appositive phrases – serve to perform a descriptive function, and, importantly, are all unnecessary for distinction! In all of these cases, the author is using the referring expressions as a vehicle for communicating information about the referents. This descriptive information is sometimes new, sometimes necessary for understanding the text, and sometimes just for added flavor. But when the expression is *descriptive*, as opposed to *distinctive*, this additional information is not required for identifying the referent of the expression, and it is these sorts of referring expressions that we will be concerned with here.

Although these sorts of referring expression have been mostly ignored by researchers in this area[1], we show in this corpus study that descriptive expressions are in fact quite prevalent: nearly one-fifth of referring expressions in news and narrative are descriptive. In particular, our data, the trained judgments of native English speakers, show that 18% of all distinctive referring expressions in news and 17% of those in narrative folktales are descriptive. With this as motivation, we argue that descriptive referring expressions must be studied more carefully, especially as the field progresses from referring in a physical, immediate context (like that in the REG Challenges) to generating more literary forms of text.

2 Corpus Annotation

This is a corpus study; our procedure was therefore to define our annotation guidelines (Section 2.1), select texts to annotate (2.2), create an annotation tool for our annotators (2.3), and, finally, train annotators, have them annotate referring expressions' constituents and function, and then adjudicate the double-annotated texts into a gold standard (2.4).

2.1 Definitions

We wrote an annotation guide explaining the difference between distinctive and descriptive referring expressions. We used the guide when training annotators, and it was available to them while annotating. With limited space here we can only give an outline of what is contained in the guide; for full details see (Finlayson and Hervás, 2010a).

Referring Expressions We defined referring expressions as referential noun phrases and their coreferential expressions, e.g., "John kissed Mary. She blushed.". This included referring expressions to generics (e.g., "Lions are fierce"), dates, times, and numbers, as well as events if they were referred to using a noun phrase. We included in each referring expression all the determiners, quantifiers, adjectives, appositives, and prepositional phrases that syntactically attached to that expression. When referring expressions were nested, all the nested referring expressions were also marked separately.

Nuclei vs. Modifiers In the only previous corpus study of descriptive referring expressions, on museum labels, Cheng et al. (2001) noted that descriptive information is often integrated into referring expressions using modifiers to the head noun. To study this, and to allow our results to be more closely compared with Cheng's, we had our annotators split referring expressions into their constituents, portions called either *nuclei* or *modifiers*. The nuclei were the portions of the referring expression that performed the 'core' referring function; the modifiers were those portions that could be varied, syntactically speaking, independently of the nuclei. Annotators then assigned a distinctive or descriptive function to each constituent, rather than the referring expression as a whole.

Normally, the nuclei corresponded to the head of the noun phrase. In (1), the nucleus is the token *king*, which we have here surrounded with square brackets. The modifiers, surrounded by parentheses, are *The* and *old*.

(1) *(The) (old) [king] was wise.*

Phrasal modifiers were marked as single modifiers, for example, in (2).

(2) *(The) [roof] (of the house) collapsed.*

It is significant that we had our annotators mark and tag the nuclei of referring expressions. Cheng and colleagues only mentioned the possibility that additional information could be introduced in the modifiers. However, O'Donnell et al. (1998) observed that often the choice of head noun can also influence the function of a referring expression. Consider (3), in which the word *villain* is used to refer to the King.

(3) *The King assumed the throne today.*
I don't trust (that) [villain] one bit.

The speaker could have merely used *him* to refer to the King–the choice of that particular head noun *villain* gives us additional information about the disposition of the speaker. Thus *villain* is descriptive.

Function: Distinctive vs. Descriptive As already noted, instead of tagging the whole referring expression, annotators tagged each constituent (nuclei and modifiers) as distinctive or descriptive.

The two main tests for determining descriptiveness were (a) if presence of the constituent was unnecessary for identifying the referent, or (b) if

[1] With the exception of a small amount of work, discussed in Section 4.

the constituent was expressed using unusual or ostentatious word choice. If either was true, the constituent was considered descriptive; otherwise, it was tagged as distinctive. In cases where the constituent was completely irrelevant to identifying the referent, it was tagged as descriptive. For example, in the folktale *The Princess and the Pea*, from which (1) was extracted, there is only one king in the entire story. Thus, in that story, *the king* is sufficient for identification, and therefore the modifier *old* is descriptive. This points out the importance of context in determining distinctiveness or descriptiveness; if there had been a roomful of kings, the tags on those modifiers would have been reversed.

There is some question as to whether copular predicates, such as *the plumber* in (4), are actually referring expressions.

(4) *John is the plumber*

Our annotators marked and tagged these constructions as normal referring expressions, but they added an additional flag to identify them as copular predicates. We then excluded these constructions from our final analysis. Note that copular predicates were treated differently from appositives: in appositives the predicate was included in the referring expression, and in most cases (again, depending on context) was marked descriptive (e.g., John, the plumber, slept.).

2.2 Text Selection

Our corpus comprised 62 texts, all originally written in English, from two different genres, news and folktales. We began with 30 folktales of different sizes, totaling 12,050 words. These texts were used in a previous work on the influence of dialogues on anaphora resolution algorithms (Aggarwal et al., 2009); they were assembled with an eye toward including different styles, different authors, and different time periods. Following this, we matched, approximately, the number of words in the folktales by selecting 32 texts from Wall Street Journal section of the Penn Treebank (Marcus et al., 1993). These texts were selected at random from the first 200 texts in the corpus.

2.3 The Story Workbench

We used the Story Workbench application (Finlayson, 2008) to actually perform the annotation. The Story Workbench is a semantic annotation program that, among other things, includes the ability to annotate referring expressions and coreferential relationships. We added the ability to annotate nuclei, modifiers, and their functions by writing a workbench "plugin" in Java that could be installed in the application.

The Story Workbench is not yet available to the public at large, being in a limited distribution beta testing phase. The developers plan to release it as free software within the next year. At that time, we also plan to release our plugin as free, downloadable software.

2.4 Annotation & Adjudication

The main task of the study was the annotation of the constituents of each referring expression, as well as the function (distinctive or descriptive) of each constituent. The system generated a first pass of constituent analysis, but did not mark functions. We hired two native English annotators, neither of whom had any linguistics background, who corrected these automatically-generated constituent analyses, and tagged each constituent as descriptive or distinctive. Every text was annotated by both annotators. Adjudication of the differences was conducted by discussion between the two annotators; the second author moderated these discussions and settled irreconcilable disagreements. We followed a "train-as-you-go" paradigm, where there was no distinct training period, but rather adjudication proceeded in step with annotation, and annotators received feedback during those sessions.

We calculated two measures of inter-annotator agreement: a kappa statistic and an f-measure, shown in Table 1. All of our f-measures indicated that annotators agreed almost perfectly on the location of referring expressions and their breakdown into constituents. These agreement calculations were performed on the annotators' original corrected texts.

All the kappa statistics were calculated for two tags (nuclei vs. modifier for the constituents, and distinctive vs. descriptive for the functions) over both each token assigned to a nucleus or modifier and each referring expression pair. Our kappas indicate moderate to good agreement, especially for the folktales. These results are expected because of the inherent subjectivity of language. During the adjudication sessions it became clear that different people do not consider the same information

as obvious or descriptive for the same concepts, and even the contexts deduced by each annotators from the texts were sometimes substantially different.

	Tales	Articles	Total
Ref. Exp. (F_1)	1.00	0.99	0.99
Constituents (F_1)	0.99	0.98	0.98
Nuc./Mod. (κ)	0.97	0.95	0.96
Const. Func. (κ)	0.61	0.48	0.54
Ref. Exp. Func. (κ)	0.65	0.54	**0.59**

Table 1: Inter-annotator agreement measures

3 Results

Table 2 lists the primary results of the study. We considered a referring expression descriptive if any of its constituents were descriptive. Thus, 18% of the referring expressions in the corpus added additional information beyond what was required to unambiguously identify their referent. The results were similar in both genres.

	Tales	Articles	Total
Texts	30	32	62
Words	12,050	12,372	24,422
Sentences	904	571	1,475
Ref. Exp.	3,681	3,526	7,207
Dist. Ref. Exp.	3,057	2,830	5,887
Desc. Ref. Exp.	609	672	1,281
% Dist. Ref.	83%	81%	82%
% Desc. Ref.	17%	19%	**18%**

Table 2: Primary results.

Table 3 contains the percentages of descriptive and distinctive tags broken down by constituent. Like Cheng's results, our analysis shows that descriptive referring expressions make up a significant fraction of all referring expressions. Although Cheng did not examine nuclei, our results show that the use of descriptive nuclei is small but not negligible.

4 Relation to the Field

Researchers working on generating referring expressions typically acknowledge that referring expressions can perform functions other than distinction. Despite this widespread acknowledgment, researchers have, for the most part, explicitly ignored these functions. Exceptions to this trend

	Tales	Articles	Total
Nuclei	3,666	3,502	7,168
Max. Nuc/Ref	1	1	1
Dist. Nuc.	95%	97%	96%
Desc. Nuc.	5%	3%	**4%**
Modifiers	2,277	3,627	5,904
Avg. Mod/Ref	0.6	1.0	0.8
Max. Mod/Ref	4	6	6
Dist. Mod.	78%	81%	80%
Desc. Mod.	22%	19%	**20%**

Table 3: Breakdown of Constituent Tags

are three. First is the general study of *aggregation* in the process of referring expression generation. Second and third are corpus studies by Cheng et al. (2001) and Jordan (2000a) that bear on the prevalence of descriptive referring expressions.

The NLG subtask of aggregation can be used to imbue referring expressions with a descriptive function (Reiter and Dale, 2000, §5.3). There is a specific kind of aggregation called *embedding* that moves information from one clause to another inside the structure of a separate noun phrase. This type of aggregation can be used to transform two sentences such as *"The princess lived in a castle. She was pretty"* into *"The pretty princess lived in a castle"*. The adjective *pretty*, previously a copular predicate, becomes a descriptive modifier of the reference to the princess, making the second text more natural and fluent. This kind of aggregation is widely used by humans for making the discourse more compact and efficient. In order to create NLG systems with this ability, we must take into account the caveat, noted by Cheng (1998), that any non-distinctive information in a referring expression must not lead to confusion about the distinctive function of the referring expression. This is by no means a trivial problem – this sort of aggregation interferes with referring and coherence planning at both a local and global level (Cheng and Mellish, 2000; Cheng et al., 2001). It is clear, from the current state of the art of NLG, that we have not yet obtained a deep enough understanding of aggregation to enable us to handle these interactions. More research on the topic is needed.

Two previous corpus studies have looked at the use of descriptive referring expressions. The first showed explicitly that people craft descriptive referring expressions to accomplish different

goals. Jordan and colleagues (Jordan, 2000b; Jordan, 2000a) examined the use of referring expressions using the COCONUT corpus (Eugenio et al., 1998). They tested how domain and discourse goals can influence the content of non-pronominal referring expressions in a dialogue context, checking whether or not a subject's goals led them to include non-referring information in a referring expression. Their results are intriguing because they point toward heretofore unexamined constraints, utilities and expectations (possibly genre- or style-dependent) that may underlie the use of descriptive information to perform different functions, and are not yet captured by aggregation modules in particular or NLG systems in general.

In the other corpus study, which partially inspired this work, Cheng and colleagues analyzed a set of museum descriptions, the GNOME corpus (Poesio, 2004), for the pragmatic functions of referring expressions. They had three functions in their study, in contrast to our two. Their first function (marked by their `uniq` tag) was equivalent to our distinctive function. The other two were specializations of our descriptive tag, where they differentiated between additional information that helped to understand the text (`int`), or additional information not necessary for understanding (`attr`). Despite their annotators seeming to have trouble distinguishing between the latter two tags, they did achieve good overall inter-annotator agreement. They identified 1,863 modifiers to referring expressions in their corpus, of which 47.3% fulfilled a descriptive (`attr` or `int`) function. This is supportive of our main assertion, namely, that descriptive referring expressions, not only crucial for efficient and fluent text, are actually a significant phenomenon. It is interesting, though, that Cheng's fraction of descriptive referring expression was so much higher than ours (47.3% versus our 18%). We attribute this substantial difference to genre, in that Cheng studied museum labels, in which the writer is space-constrained, having to pack a lot of information into a small label. The issue bears further study, and perhaps will lead to insights into differences in writing style that may be attributed to author or genre.

5 Contributions

We make two contributions in this paper.

First, we assembled, double-annotated, and adjudicated into a gold-standard a corpus of 24,422 words. We marked all referring expressions, coreferential relations, and referring expression constituents, and tagged each constituent as having a descriptive or distinctive function. We wrote an annotation guide and created software that allows the annotation of this information in free text. The corpus and the guide are available on-line in a permanent digital archive (Finlayson and Hervás, 2010a; Finlayson and Hervás, 2010b). The software will also be released in the same archive when the Story Workbench annotation application is released to the public. This corpus will be useful for the automatic generation and analysis of both descriptive and distinctive referring expressions. Any kind of system intended to generate text as humans do must take into account that identification is not the only function of referring expressions. Many analysis applications would benefit from the automatic recognition of descriptive referring expressions.

Second, we demonstrated that descriptive referring expressions comprise a substantial fraction (18%) of the referring expressions in news and narrative. Along with museum descriptions, studied by Cheng, it seems that news and narrative are genres where authors naturally use a large number of descriptive referring expressions. Given that so little work has been done on descriptive referring expressions, this indicates that the field would be well served by focusing more attention on this phenomenon.

Acknowledgments

This work was supported in part by the Air Force Office of Scientific Research under grant number A9550-05-1-0321, as well as by the Office of Naval Research under award number N00014091059. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Office of Naval Research. This research is also partially funded the Spanish Ministry of Education and Science (TIN2009-14659-C03-01) and Universidad Complutense de Madrid (GR58/08). We also thank Whitman Richards, Ozlem Uzuner, Peter Szolovits, Patrick Winston, Pablo Gervás, and Mark Seifter for their helpful comments and discussion, and thank our annotators Saam Batmanghelidj and Geneva Trotter.

References

Alaukik Aggarwal, Pablo Gervás, and Raquel Hervás. 2009. Measuring the influence of errors induced by the presence of dialogues in reference clustering of narrative text. In *Proceedings of ICON-2009: 7th International Conference on Natural Language Processing*, India. Macmillan Publishers.

Douglas E. Appelt. 1985. Planning English referring expressions. *Artificial Intelligence*, 26:1–33.

Hua Cheng and Chris Mellish. 2000. Capturing the interaction between aggregation and text planning in two generation systems. In *INLG '00: First international conference on Natural Language Generation 2000*, pages 186–193, Morristown, NJ, USA. Association for Computational Linguistics.

Hua Cheng, Massimo Poesio, Renate Henschel, and Chris Mellish. 2001. Corpus-based np modifier generation. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Hua Cheng. 1998. Embedding new information into referring expressions. In *ACL-36: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1478–1480, Morristown, NJ, USA. Association for Computational Linguistics.

Barbara Di Eugenio, Johanna D. Moore, Pamela W. Jordan, and Richmond H. Thomason. 1998. An empirical investigation of proposals in collaborative dialogues. In *Proceedings of the 17th international conference on Computational linguistics*, pages 325–329, Morristown, NJ, USA. Association for Computational Linguistics.

Mark A. Finlayson and Raquel Hervás. 2010a. Annotation guide for the UCM/MIT indications, referring expressions, and coreference corpus (UMIREC corpus). Technical Report MIT-CSAIL-TR-2010-025, MIT Computer Science and Artificial Intelligence Laboratory. http://hdl.handle.net/1721.1/54765.

Mark A. Finlayson and Raquel Hervás. 2010b. UCM/MIT indications, referring expressions, and coreference corpus (UMIREC corpus). Work product, MIT Computer Science and Artificial Intelligence Laboratory. http://hdl.handle.net/1721.1/54766.

Mark A. Finlayson. 2008. Collecting semantics in the wild: The Story Workbench. In *Proceedings of the AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence*, pages 46–53, Menlo Park, CA, USA. AAAI Press.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA-REG challenge 2009: overview and evaluation results. In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 174–182, Morristown, NJ, USA. Association for Computational Linguistics.

Pamela W. Jordan. 2000a. Can nominal expressions achieve multiple goals?: an empirical study. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 142–149, Morristown, NJ, USA. Association for Computational Linguistics.

Pamela W. Jordan. 2000b. Influences on attribute selection in redescriptions: A corpus study. In *Proceedings of CogSci2000*, pages 250–255.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Michael O'Donnell, Hua Cheng, and Janet Hitzeman. 1998. Integrating referring and informing in NP planning. In *Proceedings of COLING-ACL'98 Workshop on the Computational Treatment of Nominals*, pages 46–56.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the GNOME corpus. In *DiscAnnotation '04: Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79, Morristown, NJ, USA. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1992. A fast algorithm for the generation of referring expressions. In *Proceedings of the 14th conference on Computational linguistics*, Nantes, France.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. 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.*

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expressions. In *Proceedings of the 5th International Conference on Natural Language Generation.*