Determiner-Established Deixis to Communicative Artifacts in Pedagogical Text

Shomir Wilson[1,2] **and Jon Oberlander**[1]
[1]School of Informatics, University of Edinburgh, United Kingdom
[2]School of Computer Science, Carnegie Mellon University, USA
shomir@cs.cmu.edu, jon@inf.ed.ac.uk

Abstract

Pedagogical materials frequently contain deixis to communicative artifacts such as textual structures (e.g., sections and lists), discourse entities, and illustrations. By relating such artifacts to the prose, deixis plays an essential role in structuring the flow of information in informative writing. However, existing language technologies have largely overlooked this mechanism. We examine properties of deixis to communicative artifacts using a corpus rich in determiner-established instances of the phenomenon (e.g., "this section", "these equations", "those reasons") from Wikibooks, a collection of learning texts. We use this corpus in combination with WordNet to determine a set of word senses that are characteristic of the phenomenon, showing its diversity and validating intuitions about its qualities. The results motivate further research to extract the connections encoded by such deixis, with the goals of enhancing tools to present pedagogical e-texts to readers and, more broadly, improving language technologies that rely on deictic phenomena.

1 Introduction

Deixis often appears in written language as an anaphoric mechanism to refer to communicative entities in a document. Such deixis can have a variety of referent types. For example, consider *that idea* in Sentence (1), *those names* in (2), *this section* in (3), and *these figures* in (4):

(1) That idea has been challenged by many.
(2) Those names are Welsh in origin.
(3) In this section, we cover some early work.
(4) Quantities in these figures are approximate.

The kinds of deixis represented in (1) and (2) are similar to discourse deixis (Webber, 1991) and textual deixis (Lyons, 1977), respectively. Sentence (3) contains deixis to a structural element of a document (Paraboni and Deemter, 2006), and (4) contains an example of deixis to illustrative items such as figures or examples. We collectively term such deictic acts as *communicative deixis* (*CD* for brevity), recognizing their shared characteristics, and we name their referents *communicative artifacts* (*CA*s). Prior studies have focused on narrow varieties of CD (such as those identified above), leaving unknown their properties when viewed together as a whole. Moreover, efforts to automatically identify or resolve CD have been piecemeal at best. Given the complexity of the referents, conventional tools for coreference or anaphora resolution are poorly applicable.

This paper describes analysis of the first collection of instances of deixis in English targeted to refer to a broad variety of CAs. Texts from the website *Wikibooks* are used, for the intuitive density of CD in pedagogical material and the potential value of augmenting them with interpretive metadata. The diversity of referents in this corpus enables new inferences on the composition and relative frequencies of CD varieties in text. We focus on *determiner-established* instances, i.e., anaphoric noun phrases that begin with determiners *this*, *that*, *these*, or *those* (e.g., (1)-(4)). This focus has the advantage of collecting instances that explicitly identify the relevant capacities of their referents (e.g., (1) reifies its referent as an "idea").

The remainder of this paper is structured as follows. Section 2 surveys related work on deixis to specific types of CAs. Section 3 describes the text source for this study and the procedure used to collect and label instances. Section 4 describes our use of WordNet to characterize CAs, resulting in an ontology of such referents and inter-annotator agreement results for labeling of artifact types. Finally, Section 5 provides some conclusions and directions for future work.

2 Related Work

The value of CD in pedagogical contexts has been established by studies such as those by Mayer (2009) and Buisine and Martin (2007). Those motivate our work to fill the present lack of corpus-based linguistic knowledge of the phenomenon. Also, although spatial deixis falls beyond the scope of this paper, we acknowledge the efforts of others such as Gergle et al. (2013) to study its value in collaborative communication.

Prior works have examined *discourse deixis* in text, though little attention has been given to CD as a phenomenon or deixis to other CAs. Seminal papers by Webber (1988, 1991) established the importance of discourse deixis, although they focused upon demonstrative pronouns such as "this" or "that". Many efforts have addressed discourse deixis in the context of anaphora; these include Poesio and Artstein's (2008), who created a corpus of anaphoric relations inclusive of (but not limited to) discourse. Their collection included 455 instances of discourse deixis, although they noted ambiguity in the set of markables. Dipper and Zinsmeister (2012) also addressed discourse deixis through anaphora resolution and produced a collection of 225 abstract anaphors out of 643 candidate instances.

Prior studies of *shell nouns* revealed capacities of referents similar to a subset of those found in our work. Such nouns are used anaphorically to refer to complex, proposition-like pieces of information such as points, assumptions, or acts (Schmid, 2000). Kolhatkar et al. (2013) noted the pervasiveness of shell nouns in text and their tendency to "characterize and label" their antecedents. However, such antecedents only partly intersect with CAs. The set of shell nouns studied by Schmid did not include typical document entities such as *section*, *figure*, or *list*. Simultaneously, the set included many nouns with little or no relevance as CAs, such as *fury*, *miracle*, and *pride*.

The task of identifying CD in text and referent CAs bears some similarity to coreference resolution. However, coreference resolvers tried by the authors (namely CoreNLP (Recasens et al., 2013), ArkRef (O'Connor and Heilman, 2013) and the work of Roth and Bengston (2008)) were ineffective at this task. We posit that many CAs are not noun phrases, which makes them difficult or inappropriate to characterize as referring expressions. This limits the effectiveness of traditional approaches to coreference resolution toward the present problem.

Statistic	Total	Min.	Median	Mean	Max.
Words	2883178	1721	20337	23633	57465
Sentences	114474	71	832	938	2121
Candidates	10495	4	85	86	285

Table 1. Statistics for the 122 selected printable Wikibooks and the candidate instances of CD.

Our results are further distinct from prior work by focusing on the communicative capacities of a variety of referents represented in documents. However, the present focus upon determiner-established phrases is more exclusive, and our results do not include demarcation of referents. We posit that the tradeoff is worthwhile, given limited prior work on identifying CD and the lack of prior efforts to study CAs other than discourse entities.

3 Corpus Creation

Textbooks from *Wikibooks* were chosen to supply pedagogical text. Among the alternatives, this source provided the largest volume of material with a license amenable to corpus redistribution. Moreover, the collection of English language textbooks on the site covers a diverse set of topics and contains samples from a variety of writers. Below we describe our text pre-processing and then explain how candidate instances of CD were identified.

3.1 Source Material

To simplify collection and processing, 122 Wikibooks textbooks with printable versions were selected for use. Contained in this set are textbooks in eleven different subject areas, such as computing, humanities, and the sciences. In preparation for analysis, the documents were POS tagged and parsed by the Stanford CoreNLP suite (Socher et al., 2013; Toutanova et al., 2003). Table 1 presents some statistics on the texts in aggregate. They illustrate the substantial size of most texts, though a few were freshly started or incomplete. Overall, the corpus is comparable in size with corpora from efforts cited in Section 2, though text genera and sought markables vary.

Next, potential instances of CD were identified. Such instances were noun phrases beginning with determiners *this*, *that*, *these*, or *those*. We include *these* and *those* to collect CD to sets of entities, a nuance absent from any previous work. 9252 sentences, or 8% of the corpus, contained at least one potential instance.

Lemma	Freq.	Lemma	Freq.
page	314	function	83
book	287	chapter	73
case	249	information	70
example	126	problem	69
point	121	value	62
section	116	type	59
way	112	process	56
option	102	feature	56
time	101	number	54
message	93	text	54

Table 2. The 20 most frequent head nouns in candidate instances.

This collection contained substantial boilerplate text, and sentences that appeared verbatim in at least ten different books were discarded. This filtering produced a set of 7613 *candidate instances*. Table 2 shows the most frequent head nouns in candidate instances. Some resemble the shell nouns of prior work, but the presence of others illustrates the diversity of CD. Diversity was expected from pedagogical texts and validates Wikibooks as a rich source of CD.

We conducted a preliminary survey of the corpus contents by reading a random selection of 10% of candidates and judging their statuses as instances of CA. Table 3 shows examples of candidate instances, categorized by the foci of prior studies (cited in the Introduction) of CD phenomena. The researchers estimated that 48% of candidates were instances of CD, although directly labeling large numbers of candidates was deemed impractical. Instead, we noted that the word sense of the noun in a candidate instance is an important (albeit not definitive) indication of its CD status. Accordingly, we shift our focus from individual candidate instances to words that appear in them (i.e., lemmas) and word senses.

```
For each synset gloss, perform the
following:

Imagine instantiating the type
represented by the gloss. Judge its
suitability for the following statements.

(1) [an instantiation of the type] is
about a topic.

(2) [an instantiation of the type] is
intended to communicate an idea.

(3) [an instantiation of the type] can
be produced in a document or as a
document to convey information.

If at least two of the three statements
above are coherent, mark 'y' for the
gloss. Otherwise, mark 'n'.
```

Figure 1. Instructions given to annotators.

3.2 Word Senses

The noun in an instance of CD has a doubly salient role in CA, by providing a cue to the intended referent and also by reifying the referent. For example, an illustrating referent might be referred to as “this example” or “this ideal”, with divergent consequences. The noun choice semantically identifies the relevant capacity of the referent, affecting its message.

To identify the varieties and characteristics of CD in pedagogical text, we examine in aggregate the senses of those words that appear in candidate phrases in the corpus. WordNet 3.0 (Fellbaum, 1998) was chosen to provide an ontological structure for relevant word senses and thus for CAs. First, synsets for the 27 most frequent nouns in candidate phrases were collected, irrespective of viability for CD. This covered 34% of candidate instances and resulted in a set of 200 synsets. Their glosses were labeled as viable or non-viable for CD by two expert annotators, who first worked separately and then collaborated to resolve differences in their annotations.

Category	Examples
Structural	Many of the resources listed elsewhere in **this section** have…
	In **this chapter**, we will show you how to draw…
Illustrative	Consider **these sentences**: [followed by example sentences]
	[following a source code fragment] …the first time the computer sees **this statement**, ‘a’ is zero, so it is less than 10.
Discourse	Utilizing **this idea**, subunit analogies were invented…
	In **this case**, you’ve narrowed the topic down to “Badges.”
Non-CD	Devices similar to resistors turn **this energy** into light, motion…
	What type of things does a person in **that career field** know?

Table 3. Examples of candidate instances. Bold text denotes the determiner and head noun in each instance. Sentences are truncated in the table for brevity.

Figure 1 shows the annotation instructions, which were designed to address the combined range of CAs from prior work. To illustrate its application, consider the noun *chapter*. One gloss of *chapter* is “a subdivision of a written work; usually numbered and titled”. This sense clearly satisfies the third numbered statement in Figure 1. Coherency arguments for the first and second statements are less definitional, but both annotators decided at least one was satisfactory, leading to a *y* mark. Another gloss of *chapter* is “any distinct period in history or in a person’s life”. This sense fails to satisfy the second or third statement, leading to an *n* mark.

4 Results and Discussion

Resolving differences between the annotators’ labels produced a set of 62 synsets whose glosses characterized CAs. We refer to the sets of 200 synsets and 62 synsets as the CCS (candidates for communicative senses) and VCS (verified communicative senses) sets, respectively. We offer the complete results of our annotations online[1] to encourage further research on this topic. In this section we present inter-annotator agreement statistics and describe the composition of the VCS set using the structure of WordNet.

4.1 Inter-Annotator Agreement

The kappa statistic for category agreement between the two annotators was 0.70, with matching annotations on 174 of 200 senses. Although this metric is an imperfect indicator, this value is generally regarded as substantial (Viera and Garrett, 2005) albeit with some tentativeness (Carletta, 1996). The annotators respectively placed 33% and 30% of instances in the VCS set, suggesting general agreement on the distribution of labels irrespective of specific instances. The annotators agreed that some cases were difficult to label without context, and a combination of sense labeling and in-text instance labeling may be fruitful for future work.

4.2 Representation in WordNet

We use the structure of WordNet to illustrate the properties of CAs that VCS senses represent. To do this, the hypernym closure (i.e., the sequence(s) of hypernyms from a given synset to the root synset) was computed for each VCS sense. These “traces” were aggregated into a reproduction of a subset of WordNet’s synsets and relations, resulting in a *de facto* ontology of CAs. The same procedure was performed for the CCS set to create an illustrative baseline.

Synset	CCS	VCS	Chg.
0 entity.n.01	217 / 217	72 / 72	0
1 abstraction.n.06	166 / 217	65 / 72	.14
2 psych._feature.n.01	51 / 166	15 / 65	-.08
2 communication.n.02	47 / 166	37 / 65	.29
2 attribute.n.02	24 / 166	2 / 65	-.11
2 group.n.01	18 / 166	4 / 65	-.05
2 measure.n.02	15 / 166	3 / 65	-.04
2 relation.n.01	11 / 166	4 / 65	.00
1 physical_entity.n.01	51 / 217	7 / 72	-.14
2 object.n.01	38 / 51	6 / 7	.11
2 causal_agent.n.01	7 / 51	0 / 7	-.14
2 thing.n.12	4 / 51	0 / 7	-.08
2 process.n.06	1 / 51	0 / 7	-.02
2 matter.n.03	1 / 51	1 / 7	.12

Table 4. Distributions of traces through the first two hyponym relations emanating from the root synset *entity.n.01*, for CCS and VCS. Fractions indicate the constituent weight of each synset.

Table 4 shows the structure of the most general synsets in the ontologies constructed from VCS and CCS traces. Fractions illustrate the relative constituent weight of each synset, by virtue of the traces that include it. For example, 65 of the 72 traces for VCS synsets pass through *abstraction.n.06*, and 37 of those 65 traces pass through *communication.n.02*. The total quantities of traces for CCS and VCS are greater than their respective set sizes because of a small number of synsets in those sets with multiple hyponym paths to the root. The rightmost column of Table 4 shows the decimal result of subtracting the CCS constituent weight fraction from the VCS fraction. Positive numbers indicate that the manual labeling of senses magnified the weight of a synset over the CCS baseline.

The constituent weights confirm some intuitions but also hold a few surprises. The vast majority of CAs are abstractions rather than physical entities, and most of the abstractions are “something that is communicated by or to or between people or groups” (the gloss of *communication.n.02*). Psychological features are also a substantial constituency, with traces to VCS synsets that represent words such as *method*, *plan*, and *question*. Most of the few VCS physical entities are communicative artifacts in their complete form (e.g., a book or a periodical issue). *Matter* as a physical entity may seem out of place in Table 4. The VCS synset responsible for its inclusion is *page.n.01*, which

[1] `http://www.cs.cmu.edu/~shomir/wb_cd_study/`

has the gloss "one side of one leaf (of a book or magazine or newspaper or letter etc.) or the written or pictorial matter it contains." Both annotators believed it merited inclusion in VCS.

Finally, we observed that many VCS senses (58%) were not the first sense for their words, indicating different senses appear more often[2]. This likely hinders word sense disambiguation of nouns in CD instances: the common baseline of first sense tagging is futile in these cases, and their extra-topical nature means that appropriate CA senses are not implied by the surrounding words (Wilson, 2011). This suggests that identification of CD instances may require a dedicated approach to word sense tagging.

5 Conclusion

The results of this study illustrate the significance of CD, both for the processing of pedagogical texts and for the broader project of understanding anaphora. Its pervasiveness and its diversity show its potential as a conduit for language technologies to enrich documents with pragmatic metadata. Our next effort will be to identify the referents of CD instances using knowledge from the present study of the character and distribution of those referents. CAs are represented by spans of content in a document (e.g., text or figures), and accordingly the identification of a CD referent will involve the selection of the correct span of content. We expect that the word sense of the noun in a CD phrase will limit the set of potentially relevant CAs, and that both localized features (such as paragraph position of a CD instance and the expected CA count) and document-level features (e.g., proximity of potential referents) will be valuable.

Acknowledgment

This research was supported by grant #1159236 from the US National Science Foundation's International Research Fellowship Program.

References

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proc. EMNLP*.

Buisine, S. and Martin, J.-C. (2007). The effects of speech–gesture cooperation in animated agents' behavior in multimedia presentations. *Interacting with Computers*, *19*(4), 484–493. doi:10.1016/j.intcom.2007.04.002

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22*(2), 249–254.

Dipper, S. and Zinsmeister, H. (2012). Annotating abstract anaphora. In *Proc. LREC*, *46*(1), 37–52. doi:10.1007/s10579-011-9160-1

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Gergle, D., Kraut, R. E., and Fussell, S. R. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human-Computer Interaction*, *28*(1), 1–39.

Kolhatkar, V., Zinsmeister, H., and Hirst, G. (2013). Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proc. EMNLP* (pp. 300–310).

Lyons, J. (1977). *Semantics*. Cambridge University Press.

Mayer, R. E. (2009). *Multimedia Learning*. Cambridge University Press.

O'Connor, B. and Heilman, M. (2013). ARKref: A rule-based coreference resolution system. arXiv:1310.1975,

Paraboni, I. and Deemter, K. (2006). Referring via document parts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 3878, pp. 299–310). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/11671299_31

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU Corpus. In *Proc. LREC*. Marrakech, Morocco: European Language Resources Association (ELRA).

Recasens, M., Catherine de Marneffe, M., and Potts, C. (2013). The life and death of discourse entities: Identifying singleton mentions. In *Proc. NAACL*.

Schmid, H.-J. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Walter de Gruyter.

[2] The WordNet manual advises that senses are "generally" ordered by frequency.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013). Parsing with compositional vector grammars. In *Proc. ACL* (pp. 455–465).

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. NAACL*. doi:10.3115/1073445.1073478

Viera, A. J., and Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *Proc. ACL* (pp. 113–122). doi:10.3115/982023.982037

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. In *Natural Language and Cognitive Processes*.

Wilson, S. (2011). *A Computational Theory of the Use-Mention Distinction in Natural Language*. University of Maryland at College Park. PhD Thesis, College Park, MD, USA.