An Analysis of Anaphora Resolution in Common Botswana Newspaper Articles

Alfred J. Matiki, **University of Botswana**

Abstract

This paper examines the resolution of anaphoric expressions in newspaper articles. More specifically, the paper attempts to determine the search space for antecedents for various anaphor forms and most referent types. Using articles from popular Botswana newspapers, the study demonstrates that writers have available to them a whole array of anaphora resolution strategies which include morphologically marked number, gender, and animacy; semantic roles; grammatical relations; number of intervening referents between anaphors and their antecedents; repetition among others.

Keywords: anaphora, newspaper articles,

Introduction

Anaphora Resolution (AR) is an area of research that has attracted quite considerable interest among linguists as well as scholars in many sub-fields of natural language processing. It is an important scholarly area because natural language understanding in general is dependent on it. Similarly, machine translation, information extraction, text summarization, and information retrieval cannot be fully successful without an understanding of what such frequent words as pronouns and demonstratives refer to.

In spite of numerous studies, AR is far from being fully understood. Part of the reason for this is that any resolution of anaphora requires a number of variables to be taken into consideration. As Lord and Dahlgren (1997: 323-324) note, a comprehensive account of anaphora would require information on all anaphor forms (including pronoun, demonstrative, and full noun phrase), for all referent categories (objects, events, states, and propositions, for example), in a wide range of languages, in all discourse modes and types, ranging from spontaneous spoken language to formal speeches, novels, newspaper articles, and recipes, to name only a sampling.

An individual entity in an expository text is first introduced with an indefinite noun phrase or a proper noun. The noun phrase has a referential connection to an entity in the world. This entity is called a *referent.* A subsequent noun phrase in the text can refer to the same entity. This subsequent mention is called an *anaphor* and takes the form of a pronoun, demonstrative pronoun, noun phrase with a demonstrative adjective, or a noun phrase with a definite article or a possessive modifier. The previous mention of the referent is called the *antecedent* of the anaphor. It is possible for a text to contain several noun phrases referring to the same entity and thus creating a chain of reference. It is partly this phenomena that give texts their cohesion. Thus, grammatical form, grammatical function, agreement constraints, collocation patterns all have a bearing on anaphora resolution.

This paper contributes to the current debate on anaphora resolution. It surveys anaphora phenomena in newspaper articles published in English in Botswana and relates these anaphora patterns to anaphora as a general linguistic category. Many corpus-based studies of anaphora have been based on English in contexts where the language is native. The current study focuses on English in a non-native context in the hope that anaphora patterns that are specific to non-native contexts may be uncovered. Additionally, the focus will be on newspaper articles given that strategies of anaphora resolution are dependent on the genre and style of the text. Specifically, the study is concerned with the following factors:

- distance between anaphors and their antecedents
- forms of anaphors
- types of referents, and
- the role of global topics in anaphora resolution.

Overview of Research on Anaphora Resolution

Numerous scholars have examined the phenomena of anaphora focusing on many of its aspects. Some of these have focused only on pronominal anaphors (Hobbs 1976, Fox 1987); others have examined demonstratives (Webber 1988, Ariel 1990); few studies have looked at full noun phrase anaphors. Studies have also varied in terms of the languages studied, with English predominating (see Givón 1983), as well as with respect to the discourse mode or genre, ranging from spoken texts (Clancy 1980, Givón 1983) to newspaper articles (Lord and Dahlgren 1997, Hinds 1977). Chafe (1994), Ariel (1990, 2001), Gundel et al. (1993) and others have offered cognitive accounts for discourse anaphora and reference in general, while Givón (1983) and Levinson (2000) proposed pragmatic accounts instead.

One of the more comprehensive studies of anaphors was carried out by Lord and Dahlgren (1997) on the *Wall Street Journal* newspaper. They attempted to assess whether there is a uniform constraint on the ‘search space’ for antecedents for all anaphor forms and all referent types and whether anaphora constraints are affected by genre. They concluded that the patterns of anaphora they studied were consistent with a psychological model in which pronoun reference is limited to items in short-term memory. They also noted the importance of discourse segments as the focal point of operating memory. Items that are accessible to anaphora in general have antecedents inside the discourse segment. Anaphors with antecedents outside the discourse segment are explicit full noun phrases. Additionally, they noted that recency (anaphor-antecedent distance) is a function of discourse referent type, anaphor form as well as the type of genre.

Work by Bergler (1997) has shown that anaphora resolution at a desired level of reliability has to remain partial. The study argues that multiple small (“expert”) procedures of known reliability that are conceived for partial analysis have to be developed and combined in order to increase coverage. These resolution experts will be specific to style, domain, task, among other factors. The paper shows that over a third of co-referring NPs in the data studied were identical and could therefore be recovered reliably. A good 25% of the anaphor NPs were close to the antecedent NP in the reference chain. In general, therefore, close to 60% of co-referring NPs could be identifiable with very simple techniques. The study also showed that NP chains considered to be in the topic of an article usually require anaphora resolution and are lexically more complex than non-topical reference chains. The role of topics in anaphora resolution was also affirmed in Lord and Dahlgren's study (1997).

Other studies on anaphora have identified a number of factors which are important for anaphor form and distribution. Lord and Dahlgren (1997) note, for instance, that the relevance of number, gender, and animacy, marked morphologically on anaphors in English, is uncontroversial. Other important factors in AR include semantic role, grammatical relation, parallelism, repetition, word order, and number of intervening referents between anaphor and antecedent.

As Goecke and Witt (2006) and Strube and Müller (2003) point out, information on the possible distance between antecedent and anaphora is very crucial. Distance is usually measured in words, sentences, paragraphs, as well as discourse entities. Mitkov (2002:17) emphasizes the importance of including anaphor-antecedent distance because not only is it "interesting from the point of view of theoretical linguistics, but can be very important practically and computationally in that it can narrow down the search scope of candidates for antecedents." In spite of the importance of all these factors, a full anaphora resolution is almost impossible to attain given that anaphoric relations may be hidden in the context (Bergler 1997). Nevertheless, it will be interesting to see the extent to which these factors are adhered to by Botswana newspaper writers.

Corpus Analysis

This study is based on 20 articles, with a combined total of about 14,000 words, drawn from three newspapers published in English in Botswana. The papers included "*The Voice, Mmegi, and Botswana Guardian.*" The articles were mostly front-page stories that typically reported a news event, accompanied with background information, comments from participants, observers, or some experts on the issues. The articles were generally expository in nature.

For each article, a simple manual algorithm was used following Bergler (1997: 65), albeit with some changes:

1. Determine candidate referents within the sentence. If none are found due to lack of agreement, for instance, determine candidate referents in previous sentence.
2. Test each candidate referent for actual co-reference using:
   a) Agreement (person, number, and gender)
   b) Full NP copy (common head noun)
   c) Full NP non-copy (cognates, synonyms, nominalization)
3 If there is more than one possible co-reference, select the best.

Lord and Dahlgren's (1997) concept of *global topic (topicality)* was also crucial in resolving some anaphoric references. Global topic is a construct that is peculiar to genres like newspapers. Each newspaper article deals with an event or state which is usually identified in the article's first sentence. This is what constitutes the global topic. Each global topic statement includes noun phrases which name the relevant people and objects, the participants in the event. The event or state and the principle characters captured in the topic remain active in the reader's mind and therefore frequency and recency of mention may have correlations with the pattern of anaphora. Bergler (1997:63; also see Lundquist 1989) also uses the concept of topic, defined in this case as "one of the NPs that occur in the headline or the first sentence." Examples of global topics from the articles analyzed in this study included the following:

1. Mwakwa family puzzled by Thokolosi demands
2. Find My Son Before It's Too Late - Mother appeals for information as hunt for 13-year-old son goes into second week
3. Cocaine Hits School - Boy goes crazy after marijuana and cocaine puffs, four more suspended
4. Cabinet storm over PEEPA board

It is also important to mention that information on the logical structure of texts was also crucial in resolving anaphoric references. Such text elements as chapters, sections, and paragraphs are always useful in determining the lifespan of antecedents. As noted with respect to global topics, some linguistic expressions are more likely to be selected as antecedents throughout the whole text than others. These elements may also coincide with discourse segments which limit the search scope of candidates for antecedents and are therefore useful in anaphora resolution (see Lord and Dahlgren 1997).

Results and Discussion

This section reports and discusses the results of the corpus analysis. The range of data in terms anaphor forms, distance between anaphors and antecedents, the nature of antecedents, and referent types are described. These results are as consistent as possible for a manual analysis by a single analyst.

Anaphor forms and referent types

The study identified each anaphor and the referent type that it represented. Two referent types were identified – individual and abstract. Following Lord and Dahlgren (1997), referents in the individual category were typically people, objects, or institutions. The abstract referent types included events, states, propositions, and facts.

Table 1: Anaphor form and reference type

Reference Type	**Individual**	**Abstract**	**TOTAL**
Anaphor Form			
Pronouns	535 (97.2%)	15 (2.73%)	550 *(54.73%)*
Demonstratives	4 (40%)	6 (60%)	10 *(1%)*
Full noun phrase	315 (70.7%)	130 (29.21%)	445 *(44.28%)*
TOTAL	854 (84.98%)	151 (15.02%)	1005

As Table 1 shows, there were a total of 1005 anaphors in the corpus. Of these, 54.73% were pronouns. Demonstratives accounted for only 1%, while 44.28% were full noun phrases, with *the* or a possessive modifier. The results further show that 84.98% of the anaphors had individual referents, while the rest (15.02%) were abstract.

Like other studies have demonstrated, the anaphor forms – pronoun, demonstrative, and full noun phrases – differed from each other with respect to the likelihood that their referents were individual or abstract objects. What is significant here is the fact that any study that limits itself to any anaphor form or any referent type cannot account for a significant number of anaphoric relations.

It is also important to note from the Table 1 the preponderance of pronoun anaphors for individual referents. The correlation between full noun phrase anaphors and individual referents was the second strongest. These patterns are consistent with the kind of newspaper articles that formed the corpus for this study. All the articles studied involved events in which individual characters were central. There were stories about students being caught with drugs, a mother missing her son, politicians quarreling over the appointment board members and such other stories. This contrasts with studies that have examined, for instance, the *Wall Street Journal* (see Lord and Dahlgren 1997; Bergler 1997) in which full noun phrase anaphors as well as abstract referents are significantly higher because the articles deal, for most part, with movements in the stock market rather than individual players.

Most pronouns were nominative, “reflecting the tendency for topical information to occur early in the utterance, typically as subjects” (Lord and Dahlgren 1997:331). Without counting pronouns that had antecedents within the same sentence, the pronoun he accounted for 60.91% of the pronoun tokens, probably reflecting the fact that the majority of newsmakers are male. In the study of anaphora in the *Wall Street Journal* by Lord and Dahlgren (1997), there were no instances of *she, her,* or *hers.*

Demonstratives as we noted only accounted for 1% of the anaphors. Generally, demonstratives refer to events and most of the articles studied in this paper had individuals rather than events as the main focus. Additionally, it is reasonable to think that the writers generally avoided the demonstratives because of their inherent ambiguity in the context of multi-sentence discourse segments. Webber (1991), for instance, argues that events are only available for anaphoric reference when they are mentioned by the last utterance or by the situation that is constructed by the preceding discourse segment. Segmented Discourse Structure Theory (SDRT), on the other hand, has a different prediction. Experimental work on event anaphora resolution by Schilder (1999) has corroborated the ambiguous nature of the demonstrative *this.*

Anaphor/antecedent relationships

As we indicated above, the relationship between pronouns and their antecedents was straight forward partly because quite a good number of pronouns had antecedents within the same sentence. Even for those antecedents that were outside the sentence, such elements as agreement patterns, textual structure, and global topic made it easy to reconcile with their anaphors.

The full noun phrase and demonstrative anaphors, on the other hand, expressed their relationships with their antecedents in various ways. Some of these different linguistic forms for the antecedents are illustrated in Table 2. A number of event anaphors had the global topic as their antecedent, while others ranged from being copies of previous mention, synonyms to various other means.

Table 2: Anaphor/antecedent relationships

	Antecedent	**Anaphor**
Nominalization	(name omitted) told the appointing authority in time that he would not be available for re-appointment	his retirement
Gerundive phrase	…illegal smoking of the drug cocktail	the drug abuse spree
Infinitival complement	… decision to retire the trio from the PEEPA board	the decision
Finitive clause	he made a proposal to cabinet for the removal of finance ministry permanent secretary, (name omitted), from the PEEPA board	the proposal
Synonyms	Mother appeals for information as hunt for 13-year-old son goes into second week	the search
Copy	one of the suspended school boys	the boys
Cognate	The mother of a 13-year old boy who went missing two weeks ago, suspects that her son has been abducted for ritual purposes	the missing child
Scenario	he ended up spending a night in a police cell because he had become disorderly	the detention

Distance from antecedent

Table 3 shows that there is a high correlation between recency and anaphor form. Recency, thus distance between anaphors and their antecedents, was measured in sentence boundaries only. For purposes of this study, a sentence was defined orthographically as being bounded by capital letters and full stops (periods). The distance between an anaphor and its antecedent was calculated as the number of sentence boundaries between the anaphor and the closest previous mention of its referent.

Table 3: Anaphor Form and Distance to Antecedent (measured in Sentence Boundaries)

Distance	**1s**	**2 Ss**	**3 Ss**	**4 Ss**	**5 or more**	**TOTAL**
Anaphor form						
Pronoun	170 (54.84%)	85 (27.42%)	45 (14.52%)	10 (3.23%)	0	310
Full NP non-copy	80 (24.62%)	50 (15.38%)	30 (9.23%)	20 (6.15%)	145 (44.62%)	325
Full NP copy	15 (11.54%)	15 (11.54%)	20 (15.38%)	0	80 (61.54%)	130
TOTAL	265 (34.42%)	150 (19.48%)	95 (12.34%)	30 (3.9%)	230 (29.87%)	770

Note that the number of anaphors in Table 3 is less than that in Table 1 because the latter ignored all pronoun anaphors that had their antecedents in the same sentence. These kinds of anaphors are usually easy to resolve. Additionally, the demonstratives were analyzed as full noun phrases since all of them were made up of a demonstrative adjective followed by a noun.

The data in Table 3 shows that, in total, 34.42% of all anaphors had their antecedents in the preceding sentence but the distance between anaphor and antecedent varied with anaphor form. For instance, a little more than half (54.84%) of all the antecedents of pronouns were in the previous sentence while 61.54% of antecedents of full noun phrase copies were five or more sentences back. The number of antecedents for pronoun anaphors decreased with distance while those of full noun phrases generally increased. The pronoun anaphors that were two or more sentences away from their antecedents were all connected to their antecedents through intervening pronoun anaphors of the same kind in a chain of reference. In some cases, however, the intervening sentences were direct quotations.

The noun phrase anaphors in this table were split into two types. Full noun phrase copies were anaphors with a head noun that was a copy of the head noun of the antecedent. The full noun phrase non-copy anaphors did not contain the same head noun as their antecedent. In most cases, the head nouns of these anaphors were synonyms of various kinds to the antecedent head nouns.

Although other studies have shown that full noun phrase copies are likely to have antecedents two or more sentences back while full non-copy anaphors were likely to have antecedents in the previous sentence (see Lord and Dahlgren 1997 for instance), the current results do not support that pattern. Table 3 shows that the majority of both full copy anaphors and full non-copy anaphors occurred five or more sentences back. Except for the fact that there were more full non-copy anaphors than full copy ones, and the fact that there were slightly more antecedents in the preceding sentence for full non-copy anaphors than for full copy anaphors, their distribution in terms of distance is really indistinguishable. This pattern may reflect, for the most part, differences in writing styles rather than any peculiar behavior of the full noun phrase anaphors.

For most full non-copy anaphors, several different full noun phrases were used for the same referent. In one article, for instance, secondary school students were suspended for smoking marijuana and using cocaine. This event was subsequently referred to *as the illegal smoking of the drug cocktail, the drug abuse spree, the incident, the news,* and such a matter. This was an obvious attempt by the writer to avoid repetition. It is also worthy noting that a number of full noun phrase anaphors that were three or more sentences away had their antecedents in the global topic.

Conclusion

This study has examined anaphora resolution in newspaper articles in Botswana. Working with 20 newspaper articles, the study has shown that anaphor forms – pronoun, demonstrative, and full noun phrases – differed from each other with respect to the likelihood that their referents were individual or abstract objects. The preponderance of pronoun anaphors for individual referents was noted. The correlation between full noun phrase anaphors and individual referents was the second strongest. Other differences noted in the study pertain to the referential distance between a previous mention (antecedent) and a current mention (anaphor). There was a high correlation between recency and anaphor form. It ahs also been noted that the full noun phrase and demonstrative anaphors expressed their relationships with their antecedents in various ways. A number of event anaphors had the global topic as their antecedent, while others ranged from being copies of previous mention, synonyms to various other means. We have to agree with Fox (1987:152) that "there is no single rule for anaphora that can be specified for all of English… instead, we have a variety of specific patterns which obviously share a number of general characteristics, but which nevertheless differ enough to require separate formulations."

References

Ariel, Mira. 1990. *Accessing Noun-Phrase Antecedents.* London: Routledge.

Ariel, Mira. 2001. Accessibility theory: An overview. In Ted Sanders, Joost Schilperoord and Wilbert Spooren, eds., *Text Representation: Linguistic and Psycholinguistic Aspects.* Amsterdam: John Benjamins, 29–87.

Bergler, Sabine. 1997. Towards reliable partial anaphora resolution. *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.* Proceedings of a workshop sponsored by the Association for Computational Linguistics, Madrid, Spain.

Chafe, Wallace L. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of*

Consciousness Experience in Speaking and Writing. Chicago: University of Chicago Press.

Fox, Barbara A. 1987. The noun phrase accessibility hierarchy reinterpreted: Subject primacy or the absolutive hypothesis? *Language* 63(4):856–870.

Fox, Barbara A.1987. *Discourse Structure and Anaphora: Written and Conversational English.* Cambridge: Cambridge University Press.

Givón, Talmy. (ed). 1983. *Topic Continuity in Discourse: A Quantitative Cross-language Study.* Amsterdam: John Benjamins.

Givón, Talmy. 1983. *Topic continuity in Discourse: A Quantitative Cross-Language Study.* Amsterdam: John Benjamins.

Goecke, D. and Witt, A. 2006. Exploiting logical document structure for anaphora resolution. *Proceedings of the 5th International Conference.* Genoa, Italy.

Gundel, Jeanette K., Nancy Hedberg and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69 (2):274–307.

Hinds, John. 1977. Paragraph structure and pronominalization. *Papers in Linguistics 10*; 77-99.

Hobbs, Jerry R. 1976. Pronoun Resolution. Technical Report 76-1, Department of Computer Science, City College, City University of New York.

Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge, Mass.: MIT Press.

Lord, Carol and Kathleen Dahlgren. 1997. Participant and event anaphora in newspaper articles. In Joan Bybee, John Haiman and Sandra A Thompson, (eds.), *Essays on language Function and Language Type Dedicated to T. Givón.* Amsterdam: John Benjamins, 323-356.

Strube, M. and C. Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. ACL 03.

Webber, Bonnie Lynn. 1988. Discourse deixis: reference to discourse segments. *Proceedings of the 26th Annual meeting, Association for Computational Linguistics,* 113-122.