Towards a resolution of comparative anaphora
A corpus study of "other"

Natalia N. Modjeska
Division of Informatics
University of Edinburgh
natalian@cogsci.ed.ac.uk

Abstract

This paper reports on an first-pass annotation exercise and analysis of a corpus of examples of "other NPs" – that is, NPs modified by the comparative "other", as in "other European steelmakers". Such NPs have an anaphoric component that must be resolved – "other steelmakers than which ones?". Statistical data derived from this study have uncovered significant patterns that can be used to constrain search space and determine the correct antecedent of "other NPs". Developing efficient resolution algorithms for comparative anaphoric expression is the applied goal of this project.

1 Introduction and motivation

Comparative anaphoric expressions are noun phrases modified by, e.g., "such", "other", and "same".[1] Such phrases have two thin.gs in common: (i) they involve the notion of *comparison*, viz., identity (e.g., "same dog"), similarity (e.g., "similar dog") or dissimilarity (e.g., "different/other dog"); and (ii) they are *anaphoric*, in the sense that they rely for their interpretation on some entity or set of entities (or eventualities) in the current discourse model. (Cf. (Carter, 1987)'s definition of anaphora as

> [...] the special case of cohesion where the meaning (sense and/or reference) of one item in a cohesive relationship (the **anaphor**) is, in isolation, somehow vague and incomplete, and can only be properly interpreted by considering the meanings of the other item(s) in the relationship (the **antecedent(s)**).)

While identifying such phrases as anaphoric, it is important to distinguish them from other, more familiar cases of anaphora, such as pronominal and certain cases of nominal anaphora. Unlike pronouns and definite descriptions, "other", "such" and "similar" NPs do not require their referents to be coreferential with some other entity already in the discourse – the notable exception here are modifiers "same" and "identical", which require their referents to co-specify with their antecedents. (For an excellent recent discussion on anaphora vs. coreference see (van Deemter and Kibble, 2000).) Comparative anaphors incorporate their antecedents into their compositional semantics in the way that is idiosyncratic for each anaphor. For instance, "other", which is the focus of this paper, excludes an entity from a set of entities under consideration, providing a set-complement to something already given. E.g., in (1), the noun phrase "other European countries" refers to a set of entities characterised as "European countries", excluding "Ukraine".[2]

(1) Over four years, Ukraine would receive 75,000 million cubic metres of gas and 50,000,000-70,000,000 tons of oil, some of which would be passed on to other European countries.

[1] (Halliday and Hasan, 1976) called them *comparative reference*.

[2] This paper is not concerned with the compositional semantics of comparative expressions, and no more will be said about their contribution to the meaning of such NPs. For a detailed analysis of the semantics of comparative words see (Bierner, 2000).

In order to arrive at such interpretation, one must identify "Ukraine" as the antecedent of the "other NP" and exclude its referent from the set of European countries. (As part of the interpretation process, "Ukraine" is recognised as a European country; more about this is said in Sections 2.5 and 2.6.) The antecedent can provide a referent either *directly* or through a *mediated reference*, by linking to an entity in the discourse model – *a discourse anchor*; a discourse anchor is evoked by the antecedent. That anchor and antecedent are different entities is illustrated by the following example.

(2) She lifted the receiver as Myra darted to the other phone and, her mouth set in a straight line, dialled the number of the Roman's office.

In (2), the antecedent of "the other phone" is the NP "the receiver", but the anchor is "the telephone of which the receiver is the part" and must be derived from the antecedent. Section 2.3 examines in detail types of antecedents of "other NPs" and comments on what inference mechanisms might be involved in deriving the anchor from its linguistic source.

The work reported in this paper is part of a larger project which aims at modelling the principles by which comparative anaphoric expressions can be properly interpreted. (This may involve inference, as example (2) shows.) The applied goal of the project is to develop efficent resolution algorithms for comparative anaphoric expressions. The necessity of such algorithms is dictated by the high frequency of comparative words in written and spoken discourse. Table 1 shows the relative frequency of some comparative modifiers in the British National Corpus (BNC), a 100 million word collection of samples of written and spoken language from a variety of genres.[3] In the BNC, "other", tagged as adjective, is the 75th most common word. (There are also 35,164 occurrences of "other" tagged as noun and 14,959 occurrences tagged as pronoun.) "Such" is the 94th most common word. "Same" is on the 144th place.

Comparative modifiers	Counts	Rank
other	135,185	75
such	108,524	94
more	67,198	131
same	62,402	144
another	60,182	159

Table 1: Frequency of comparative words in the BNC

(Bierner, 2000) quotes frequency data for a number of comparative words in tutorial dialogues.[4] In 269 such dialogues, each dialogue contained, on average, 3.65 comparative phrases.[5] "Other", including the construction "X other than Y(s)", was by far the most frequent comparative word, with 686 occurrences of the total 983 comparative phrases (70%).

In view of this frequency, it is important for a variety of natural language applications that comparative words are correctly resolved. (Bierner, 2001) shows that by incorporating the analysis of comparative words into a natural language search engine, the performance of the search engine can be improved dramatically.

Comparative modifiers are also of potential interest for information extraction (IE) systems because of the "additional" knowledge that "other NPs" can bring about. For instance, in (3), an IE system capable of resolving comparative references will get a bonus in terms of a free piece of knowledge that Benidorm, like Torremolinos and Lloter, is a tourist giant – without the hassle of full text understanding.

[3]The frequency data were compiled by Adam Kilgarriff; for further details, see http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html.

[4]Bierner uses the term "alternative phrases" to refer to comparative NPs, focussing on their semantics cast in terms of *alternative sets* (Rooth, 1992).

[5]Data were collected for words "other" (including "other than"), "such", "in addition to", "besides", "another", "especially", "except", "in particular", and "unlike".

(3) But now that the resort [Benidorm] is quieter and cleaner, it more than compensates for the lack of clothes bargains. And with luck we should see the other tourist giants in Spain – Torremolinos and Lloret – moving along the same lines.

(Hearst, 1992) pioneered constructions with comparative phrases for purpose of knowledge extraction, albeit the patterns of which she gives examples are based on structural dependencies, such as constructions "$NP\ \{,NP\}*\{,\}\ and|or\ other\ NP$" and "$such\ NP\ as\ \{NP,\}*\{or|and\}\ NP$". None of the patterns that she quotes will be able to extract knowledge from the examples (1-3) above.

(Bierner, 2000) suggests a coordinated syntax-semantics approach for computing references of comparative anaphora. The technique utilises claims of the Centering Theory (Grosz et al., 1995) and is reviewed in detail in Section 3.

This paper reports on an empirical investigation into the nature of one comparative word – the modifier "other". It presents a first-pass annotation exercise and analysis of a corpus of examples of "other NPs" from the BNC along a variety of easily observable features, such as syntactic type of the antecedent, its proximity to the anaphor, and type of lexico-semantic correlation between "other" and its anchor. A similar work has been reported by (Byron and Allen, 1998) for demonstrative pronouns "this" and "that", with encouraging results. Statistical data derived from this exercise have uncovered significant patterns that can be used to constrain the search space and/or determine the correct antecedent of "other NPs". Such patterns provide a valuable complement to Bierner's resolution technique (and might boost its score, which is 47% for the BNC data). The empirical validation of these patterns is subject for future research.

The rest of the paper is organised as follows. Section 2.1 describes the corpus of the study. Section 2.2 is a meta-level description of the coding scheme which was developed for this exercise. Subsequent sections 2.3-2.5 spell out the details of the annotation and present the first findings. Section 2.6 gives examples of kinds of "additional" knowledge that can be obtained from the comparative NPs, and also reveals a difficult problem that must be resolved to make "other NPs" attractive for practical IE applications. Section 3 discusses Bierner's resolution algorithm and proposes an extension based on statistical patterns acquired from the study corpus.

2 The current study

2.1 The corpus

The BNC is a collection of written and spoken texts from a wide range of sources designed to represent a wide cross-section of current British English, both spoken and written. There are 185,308 occurrences of "other" in the BNC. Given this large number, it felt necessary to restrict the data set to a more manageable size. A 1% random sample of the BNC was used for this purpose. The sample is part of the Gsearch tool package (Keller et al., 1999), which was used to extract the examples.

The random sample contains 1,820 occurrences of the word "other" in a variety of constructions, from "other NPs", to discourse markers (e.g., "on the other hand", "in other words"), reciprocal "each other", idiomatic expressions (e.g., "the other day"), and elliptical "other" and "others".[6] In this paper, I focus on the *referential* "other NPs", i.e., those that have the potential to refer to individual objects (or sets of objects or eventualities) in the world (cf. (Fraurud, 1992)). Another condition is that "other NPs" have a full lexical head. "Other" phrases that do not fulfil these requirements, i.e., "other" as a discourse marker, idiomatic, reciprocal and elliptical uses are left out from this paper. Constructions with structurally given antecedents "X(s) other than Y(s)" are also outside the scope of this paper; in the BNC, such constructions are tagged as prepositions. Of the 445 "other" phrases in the annotated 1/3 of the sample corpus, 358 occurrences were classified as potentially referring.

[6] Elliptical "other" and "others" are tagged as nouns in the BNC, "each other" as reflexive pronouns.

2.2 Overview of the coding scheme

The annotation scheme used in this study was developed after reviewing the REFEREE co-reference annotation scheme reported in (DeCristofaro et al., 1999). The annotated features are grouped into "features of the anaphor", "features of anchor/linguistic antecedent", and "lexical relation between the anchor(s) of "other" and the class description evoked by the anaphor".

The anaphors are marked for type of determiners, pre- and postmodifiers, e.g., quantifiers, demonstrative pronouns, possessive NPs, adjectives, and numerals, as well as relative clauses and preposition phrases. (Examples are given in section 2.6.) This is "additional" linguistic information, in the sense that it is not crucially important for resolution of "other", although it is important for IE applications.

Discourse anchors of "other" can be evoked by an explicit linguistic expression, a piece of text, or they are extra-textual (situationally evoked). Antecedents, where available, are classified according to their part of speech and syntactic category, e.g., NP, adjective, proper name, or clause. Implicitly realized anchors, viz., those that do not have an explicit linguistic antecedent, were assigned a class depending on the degree of inference and type of knowledge involved. (Further details are provided in section 2.3.)

The proximity of the antecedent to the anaphor is indicated in sentence units, counting from the most recent mention. The values are "0" for conjoined contexts "X, Y and other Z(s)" (also called *list contexts*); "1" for antecedents in the same graphical sentence as the anaphor; "2" in the previous sentence; "3" two sentences (or sentence fragments) away; and "4" for antecedents further afield. Implicit anchors and cataphoric antecedents, i.e., those that follow "other NP", instead of preceding it, e.g., (4), have value "distance not marked". Discourse anchors that require inference are assigned a distance value only if some part of the text supports their derivation. (More on inference is said in sections 2.3 and 2.5.)

(4) If there is no other system of air-conditioning, and the window cannot be open all the time because of draughts or security, it should at least be opened once or twice a day for a spell.

The types of lexico-semantic relations between "other" and its anchors are described in section 2.5. The annotation reported here has not been validated by other annotators, and at presents no inter-annotator agreement score is available.

2.3 Types of anaphoric anchors and their linguistic realizations

Anchors of "other NPs" can be evoked by a variety of linguistic expressions: definite or indefinite NP[7], personal, demonstrative, or possessive pronouns, proper names, adjectives (5), clauses (6), or utterances (7).[8]

(5) EUROPEAN Community officials were stoking fears last night of an all-out trade war with the United States after it was disclosed that British and other European steelmakers could face crippling new duties on exports to America.

(6) If the patient is very heavy or the carer cannot manage for some other reason [...]

(7) How do I get my money back? Any other questions?

Multiple anchors ("split antecedent") are not unusual, e.g., (8), and can be viewed either extensionally – here, as a set containing all relevant individuals who attended the meeting, or intensionally, as "all the participants of the meeting whom I have mentioned so far".

[7] I.e., NPs with determiners "a", "an", "some", bare NPs, and cardinal plurals.
[8] The anchor in (6) is the proposition "the patient is very heavy.".

(8) Another speaker, Michael Traber, WACC's Director of Studies and Publications, said: [...] In an address on "The Cultural Environment and Media Education", Professor George Gerbner of the Annenberg School of Communication in Philadelphia also stressed [...]. The participants voted to set up an Association for Communication and Theological Education to carry forward the discussions held at Yale. The Association will be coordinated by WACC's former President, Dr William F Fore [...] Other participants at the meeting included WACC's General Secretary, Rev Carlos A Valle, and John L Peterson, Chairperson of WACC's North American Regional Association.

Frequently, "other NPs" make use of implicit anchors, relying on the speaker's and hearer's common cultural knowledge (9-10) or the utterance situation (11).

(9) Bill Clinton today prepares to stride across the political landscape as the world's most powerful man. But how does he shape up against his counterparts in other countries?

(10) The museums of the world are full of other countries' art. (The Economist, March 18, 2000, p.21)[9]

(11) LEO Jul 24-Aug 23 THERE have been three other potential points of the year when much has been primed to change — not in the easiest of ways, nor in the most predictable of fashions – and this is another one.

(9) is also an example of *metonymy*, where the hearer/reader is invited to create an inferential link from "Bill Clinton" to "the U.S.", in order to derive the anchor of "other countries".

In (11), one might think that "Jul 24-Aug 23" is the time period to be contrasted with "other potential points of the year". This is incorrect though – "Jul 24-Aug 23" is the time frame for the zodiac sign Leo. It is unclear from the example whether the horoscope was composed for "today", "this week" or "this month". The reader of the horoscope knows, though, which one is correct, from the utterance situation.

The interpretation of (12) relies on the reader's world knowledge, more specifically, the knowledge of the process of distillation. Furthermore, the NP "some other liquor" is a *metaphor*, and so is its anchor. (I leave it to the reader to decide what it might be.)

(12) The move to the stage was a logical one for Eliot, so many of whose poems have dramatic qualities. In 1920, considering "the impotence of contemporary drama", he had concluded that "The natural evolution, for us, would be to proceed in the direction indicated by Browning; to distill the dramatic essences, if we can, and infuse them into some other liquor".

The anaphoric anchor of "other" can be topic of a discourse segment (13).

(13) Well into the present century "Picklecock Alley" on Saltash Waterside was well supplied with sea-food shops. The Saltash Fair and pageant was a popular event of the 1930s, revived in the 1950s as Winkle Fair with "King Cockle". The Mayoral party would receive a formal greeting to Waterside, "which was the proper old ancient Borough centuries before the town up and over was so much as dreamed of. Fish from the Water Tamar be pretty eating as all the world knows and we offer your worships this tribute according to ancient customs of our ancestors". Pollution largely ended the shellfish industry although there have been recent attempts to revive it. Other ancient rights have been eroded away.

And finally, the anaphoric anchor or part of it might need to be derived, as examples (9) and (14) show. The inference in (9) rests on the use of proper name "Bill Clinton", the former president of the USA. In (14), the linguistic antecedent of "other pilchards" is "hogsheds that were

[9] I am grateful to Bonnie Webber for bringing this example to my attention.

Type of anchor	Counts
definite NP	84 (23%)
indefinite NP	71 (20%)
proper name	68 (19%)
common knowledge	35 (10%)
pronoun	23 (6%)
inferred	20 (6%)
proposition	14 (4%)
discourse topic	10 (3%)
utterance	8 (2%)
discourse segment	4 (1%)
demonstrative NP	2 (.5%)
adjective	2 (.5%)
undecidable	17 (5%)
total	358 (100%)

Table 2: Frequencies of anchor types in study corpus

not sold to fish merchants for export", while the anchor seems to be "pilchards that were not sold for export" and it must be derived (via inference) from "hogsheds sold for export" to "hogsheds of pilchards sold for export" to "pilchards sold for export". Information given earlier in the discourse that pilchards were salted, washed, and packed in hogsheds facilitates this inference.

(14) Now the womens' task was packing the pilchards in the "bulks" in the Cellars, laid out and salted. After twenty-eight days they would be taken out, washed and packed in hogsheds, and pressed for about ten days. Once pressed, the hogsheds, each weighting four and a quarter hundred-weight, would be sold to the fish merchants for export, mainly to Italy and other Mediterranean countries. Other pilchards were kept for home consumption.

Table 2 summarises frequencies of anchor types in the sample corpus. As evident from the table, the entity that "other NP" excludes from the set is explicitly given in more than 3/4 of cases (77%).[10] Full lexical NPs are clear leaders: definite NPs account for 23% of all anchors (including the implicit ones), indefinite NPs for 20%, and proper names for 19%. 18% of "other NPs" use implicit material as their argument: 10% require common knowledge anchors, 6% are inferred from the text, and 3% use discourse topic. It is interesting to notice that pronouns account for as little as 6% of all anchors. Psycholinguistic and computational linguistic studies have shown that pronouns are used to refer to the most salient objects. Since "other" provides a complement set to an entity already in the discourse, it is reasonable to assume that "other" with a pronominal antecedent would provide a complement set to an entity that is currently in the speaker's and hearer's center of attention. The low frequency of pronominal antecedents with "other", compared with the frequency of other antecedent types, seems to suggest that "other" is able to access less salient discourse entities. (N.B. Even if there are few pronouns in the study corpus, this doesn't mean that "other" isn't referring to the most salient or a highly salient entity. Both definite and indefinite NPs can be used for subsequent reference, and when reference is combined with predication, a definite NP is preferred over a pronoun. As Miller once noted, "There is apparently a linguistic convention that accepts anaphoric nouns that are hypernyms of the antecedent." (Miller, 1998). So, for instance, " a novel" can be subsequently referred to as "the book" as well as by "it": " I gave him a good novel, but the book/it bored him.") I will return to the issue of salience when discussing Bierner's resolution algorithm in Section 3. Demonstrative

[10]This comprises anchors evoked by definite and indefinite NPs, proper names, pronouns, clauses (giving rise to a propositional anchor), utterances, discourse segments, demonstrative NPs, and adjectives; the sum of 274 occurrences over 358 total anchors in the study corpus.

and adjectival antecedents are extremely rare in the BNC, with .5% frequencies each. Finally, in 5% of cases, it was impossible to determine the anchor.

The next section will look into the anchor distribution in relation to their proximity to the anaphor.

2.4 Proximity to the anaphorr

From the perspective of reference resolution, it is important to know not only the syntactic type of antecedent, but also where it occurs in the text. Tables 3 and 4 show the distribution of explicitly realized anchors of "other" with respect to their proximity to the anaphor. In Table 3, the spread is calculated for each antecedent type; the percentages sum up along the table rows, e.g, "24% definite NPs in list" means that of all definite NP antecedents, 24% were found in *list contexts*. List contexts, e.g., (5), are such in which the "other NP" and its antecedent occur within the same conjoined NP, joined by "and", "or", "but", "as well as", and "(along) with".[11] For details of other values see Section 2.2.

In Table 4, the focus is on list vs. non-list contexts. (The non-list value comprises "same", "previous", "two", "three" and more sentences away.) The percentages sum up for each column, indicating how likely it is for the antecedent of "other" to be realized, e.g., in a non-list context as a definite NP. For the configuration "definite" and "non-list", the probability is 33%. For cases with multiple antecedents, the distance is given for the latest antecedent only. There were no cases in the study corpus such that it wasn't possible to resolve distance.

The first phase of data analysis revealed the following three findings. (Future work might reveal other significant patterns.) First, comparative anaphora is a local phenomenon; the majority of (explicit) antecedents are found in the same sentence as "other" or the preceding one (Table 3). The figures for various types of anchors are as following: 90% for indefinite NPs, 89% for proper nouns, and 85% for definite NPs. For pronominal antecedents, the corresponding figure is 100% . Adjectival and demonstrative antecedents follow the same trend.

These figures might be compared with data for pronominal anaphors.[12] (Hobbs, 1978) reports that 98% of pronoun antecedents in his corpus were found in the same sentence as the pronoun or the previous one.

Second, as evident from Tables 3 and 4, there is a wide variation in the distribution of anchors of "other" in list vs. non-list contexts. This suggest that the two classes should be treated separately. More specifically, propositional, utterance, and discourse segment antecedents are rarely available in contexts "X, Y, and other Z(s)", as it would require all elements of the list to have a similar type of denotation, which is rare. Furthermore, the percentage of indefinite antecedents in list contexts is significantly higher than that of definite antecedents (39% and 24%, respectively), and it is more likely for a list antecedent to be realized as indefinite rather than definite NP (37% and 26%, respectively). Also, many examples of list contexts with indefinite antecedents seem to be of generic nature, as in (15).

(15) Film is capable of rising above the limitations of language and other cultural barriers.

Generic sentences express general statements about *kinds* of objects, rather than specific objects in the world. This seems to suggest that "other" may have different discourse roles in list and non-list contexts, and this is reflected in the distribution of its anchors. The exact discourse functions of "other" in these different environments remain to be worked out, but a preliminary hypothesis is that list constructions are primarily used for their classifying properties. (They are discussed in section 2.5.) Non-list "other", on the other hand, does not only characterise the anchor as belonging to a certain class, but also introduces a new referent into the discourse. (A set referent if "other Xs".)

[11] There are examples of list contexts in which the left conjunct of "other" is not its proper antecedent, e.g., "Most dogs live for about 10 years on average, and during their lives they will come into contact with possibly hundreds of people and other dogs, as well as other animals such as cats and horses." (BNC) But such examples are rare.

[12] To my knowledge, no distance figures are available for anaphoric definite NPs.

Distance						
Type of antecedent	list	same	previous	2 away	3 and more	not marked
definite NP	20 (24%)	35 (42%)	16 (19%)	2 (1.5%)	8 (10%)	4 (5%)
indefinite NP	28 (39%)	22 (31%)	14 (20%)	3 (4%)	1 (1%)	3 (4%)
proper name	25 (37%)	23 (34%)	12 (18%)	2 (3%)	6 (9%)	0
pronoun	3 (13%)	17 (77%)	2 (9%)	0	0	0
proposition	0	7 (50%)	6 (43%)	1 (7%)	0	0
utterance	0	2	3	0	3	0
discourse segment	0	1	3	0	0	0
demonstrative NP	1	0	1	0	0	0
adjective	1	1	0	0	0	0

Table 3: Proximity - variation within types with respect to distance

Type of antecedent	List	Non-list
definite NP	20 (26%)	65 (33%)
indefinite NP	28 (37%)	43 (22%)
proper name	25 (33%)	43 (22%)
pronoun	3 (4%)	19 (10%)
proposition	0	14 (7%)
utterance	0	8 (4%)
discourse segment	0	4 (2%)
demonstrative NP	1 (.5%)	1 (1%)
adjective	1 (.5%)	1 (1%)
total	76 (101%)	197 (102%)

Table 4: Characterisation of antecedent type in list and non-list contexts. In list context, this is the type of left conjunct(s)

This hypothesis is supported by the observation that (set) referents of list "other" ("other Xs" minus the anchor) are rarely referred to in the continuation of the discourse. The anchor and the set referent of "other" in list contexts are treated as a single set, and subsequent reference means reference to that set. For instance, in (16), "their" and "such victims" refer to the set of victims of drugs, including the anchor "DES daughters".[13]

(16) DES daughters and other victims of drugs would be better off if their cases were taken out of the courts. Congress could create a compensation program to help such victims while protecting the national interest in encouraging new drugs.

Subsequent mentioning of the set-referent of the "other NP" in non-list contexts, is meanwhile not uncommon (17).

(17) Absorbed in doling out "Feeding Frenzy's" tidbits, the authors gloss over the root causes of Wedtech, namely the Section 8(A) federal program under whose auspices the scandal took place. They do at least come around to saying that the courts might want to end "rigid affirmative action programs." Programs like Section 8(A) are a little like leaving gold in the street and then expressing surprise when thieves walk by to scoop it up. Numerous other scandals, among them the ones at HUD, have the same characteristics as Wedtech. They take place in government programs that seem tailor-made for corruption.

[13]The following two examples are from the Wall Street Journal. No BNC examples were available at the moment of writing.

Definite NPs are significantly more common in non-list contexts; 76% of all definite NP antecedents occur in the study corpus outside the scope of a conjoined "other NP". Furthermore, a non-list antecedent is more likely to be of a definite NP type – 33%, compared with 22% for indefinites.

Proper name antecedents do not show a clear distribution pattern, and further analysis is necessary, though the data in Table 3 suggest a slight preference for conjoined contexts. When considering proper name antecedents from a more cognitive perspective, it is important to note that proper names usually have a special cognitive status and also that it is impossible to decide from the first glance whether the name is a first mention or it has been used earlier.

The third finding concerns definite and proper name NPs in non-list contexts. A surprising 10% of all definite NP and 9% of proper name antecedents are found three and more sentences away from the anaphor. In dynamic computational models of discourse (e.g., (Strube, 1998)) the discourse model is updated with each new utterance (which by many researchers is taken to equal a sentence.) This means that discourse entities evoked by a previous sentence that are not realised in the current sentence are dropped from the list at the end of the current sentence. The update procedure reflects shifts in the center of attention of the speaker and hearer – a coherent discourse centers around one entity. With respect to "other NP", the finding that 1/10 of all definite and proper name antecedents are most recently mentioned as far as three and more sentences away from the anaphor, suggests that "other" can access referents that are no longer in the center of attention. This is perhaps not surprising, given the rich lexical content, which is available from the "other NP". Pronouns, for comparison, carry very little information and must rely on other mechanisms, such as salience.

2.5 Systematic lexical relations between anaphor and its anchor(s)

Besides contributing compositionally to the reference of "other NP", anchors of "other" stand in a systematic lexico-semantic relation with the anaphor. Two patterns have been observed by (Bierner, 2000) and (Hearst, 1992): the *instance-of* and *subclass-of* relations. The instance-of relation is a relation between an individual object and a certain class to which it is said to belong, by virtue of being the anchor of a comparative NP. In (1), for instance, the anchor "Ukraine" is an instance of the class of European countries.

A subclass-of relation holds between concepts that stand in a hypernym-hyponym relation. In (18), "a bow and arrow" is identified as a kind of weapon. In (19), the anchor "the hall" is a kind (subclass) of room.

(18) Every level has traps, baddies, bonuses and a huge nasty thing lying in wait at the end. Tiki, however, sports a handy bow and arrow and can also pick up other weapons and handy methods of transport, such as balloons, along the way.

(19) The hall is empty. There are lights in the other rooms.

(20) is a specific case of the subclass-of relation. A repeated form "benefits" is used to evoke both the anchor of "other benefits", "invalidity benefits", and its class description.

(20) People on retirement pensions have to pay tax if they have any other source of income, so why shouldn't those who receive invalidity benefit? If they are very ill they can claim other benefits, such as attendance allowance.

The third type of relation – *redescription* – is a novel one. It was first observed during the annotation exercise. Redescription is an associative relation. The class description evoked by "other" associates the anchor with a different (but compatible) class than the one to which it is known to belong. For example, in (21), "the British Clothing Industry Association", a trade organization, is identified as a sponsor of fashion shows. (21) is also an example of multiple predication; the verb "subsidise" and the predicative NP "the sole supporter" predicate properties similar to that of "sponsor". An IE system, capable of resolving anaphoric references of comparative phrases would

be able to extract this information without performing a difficult and resource-consuming task of full text understanding.

(21) Until recently the British Clothing Industry Association subsidised the event, enabling Britain's designers to show their collections in an international venue. But the association has tired of being the sole supporter and other sponsors are needed.

Re-description is available with other types of anchors than those evoked by proper names (22).

(22) This enabled the barley growers to organise themselves effectively to protest to the authorities about their loss of land, and to challenge the monopolistic price-fixing of "middlemen". The book describes the experiences of other oppressed groups in Mexico, of outcast (Dalit) communities in India, and of fisherfolk fighting for their rights in the Philippines.

Reconceptualisation of the anchor may involve metonymy (23) and metaphor (24).

(23) When the dawn came, anxious viewers on the shore could see that the waves had taken with them the Eddystone lighthouse, its eccentric architect and five other unfortunate souls.

(24) The human memory, in common with every other store, has to be positively consulted before it will function.

The systematicity of lexical relation between the anchors of "other" and the class description evoked by the anaphor is very attractive from the purpose of anaphora resolution, in particular for finding the anchor of "other". If a resolution module "knows" that the anchor of "other" belongs to a certain class, e.g., in (19), "rooms", it will need to look only for those entities that can be described as subclass of "rooms", e.g., "hall", "kitchen", or "bathroom". It does not need to consider other entities, thus reducing the search space substantially. More on resolution and potential use of lexical relations is said in section 3.

2.6 Other linguistic material in "other NPs"

"Other NPs" contain a wealth of additional linguistic material in the form of definite determiners, quantifiers (25)-(27), negation (28), and modifiers of various kinds.

(25) On this particular Friday I remembered that of course all the men were away shooting ISAAC or some other unlucky fellow.

(26) [...] it strikes me that one of the main problems with the argument is just what Ruth said: wages for housework is a kind of global undefined demand which sucks everything into it like a whirlpool, and ignores all the other demands.

(27) If the petitioner knows that the debtor has used any other names, he must state this in the petition.

(28) Persia is exceptional in the number and variety of its weaving groups. No other country can boast the same range of masterworkshop, workshop, village and nomadic rugs [...]

Whether or not "other" interacts with quantifiers, negation, and definiteness vs. indefiniteness remains to be worked out. It is clear that they affect the interpretation of the anaphor, but whether they might also affect the interpretation of its anaphoric anchor is subject for further research.

Modifiers and post-modifiers tend to supply additional information that applies to both the anaphor and its anaphoric anchor, but exceptions exist. In (29), both "Harrison", "Cornford",

and the other, unnamed, scholars are said to be classical and anthropologically influenced. In (30), on the other hand, the adjective "hemiplegic" is not applicable to the entity referred by the antecedent, because hemiplegia is the paralysis of one side of the body, and therefore only one hand is affected. In speech, this difference is usually marked by a pause after "other" and a pitch accent on "hemiplegic". In writing, such additional information is often marked by commas, but examples lacking commas, similar to (30), are not unusual.

Similarly, in (31), the relative clause "who are engaged in the struggle for justice"[14] holds for both the antecedent and its anchor. In (32), the relative clause "who could take part in debate but not vote" is exclusive of the antecedent, because "other members" are contrasted with "life peers with voting rights".

(29) Eliot moves from such a primitive organization to discussing Greek drama, following the movement of Harrison, Cornford, and the other anthropologically influenced classical scholars whom he had read.

(30) If he uses one hand on the cup handle, he should always have the other hemiplegic hand correctly positioned in front of him.

(31) In his message of congratulations, WACC's General Secretary, Rev Carlos A Valle, wrote: "We welcome this award as a recognition of your courage and commitment in the field of human rights, and we trust that it will inspire other groups and individuals who are engaged in the struggle for justice, both in Brazil and throughout Latin America."

(32) The proposal got sofar as a White Paper which suggested a two-tier system — life peers with voting rights and other members who could take part in debate but not vote.

At present, it is difficult to say what role such additional information plays in the interpretation of anaphoric expressions, and further research is necessary. It is clear, though, that such information, where applicable, must be accommodated in the discourse representations associated with the anchor(s) of "other".

3 Resolving comparative anaphora

(Bierner, 2000) suggests a coordinated syntax-semantics approach for computing references of comparative anaphora. His technique is defined within the framework of Categorial Combinatorial Grammar ((Steedman, 1996) and (Steedman, 2000)), and can be summarised as following. (For details see (Bierner, 2000).)

In list contexts, the potential antecedent(s) of "other" are found by utilising the constraints of the coordinated NP. In non-coordinated contexts, the candidates are identified via standard discourse and anaphora resolution techniques (more specifically, centering (Grosz et al., 1995) and c-command). Candidates are then subjected to a filter which filters out inappropriate candidates. This is done through presupposition binding/accommodation in the following fashion.

Lexical entities for each comparative word have a set of assertions and presuppositions associated with them. The assertional and presuppositional semantics vary slightly from one word to another, but they all presuppose that there exist an antecedent which has the properties of the anaphor (these properties are induced by the lexical head of the comparative NP) and which is an "alternative" to the anaphoric NP. When a candidate is found, it is checked against these presuppositions. If the properties of the antecedent do not match those of the anaphor, the antecedent is checked for whether its properties are consistent with the properties of the anaphor. If this is the case, the candidate is considered to be the correct antecedent, and its properties are accommodated.

If more than one candidates exist, they are considered in the ordering that roughly corresponds to grammatical role ranking $subject > object(s) > other(s)$, mediated through a salience list.

[14] I will not address the PP "both in Brazil and throughout Latin America", as its attachment is potentially ambiguous.

There are additional heuristics that help to identify the antecedents by exploiting presence of other comparative phrases in the sentence. When all possibilities have been exhausted, the algorithm suggest an extra-textual, situationally evoked antecedent, e.g., the inquirer's Web browser in the context of the question: "Where can I find web browsers for download?".

Bierner implemented a version of his algorithm, enhanced with a heuristic that chooses the most recent sentential subject as the anchor of non-list "other", and hand-tested it on three corpora: a subset of the BNC, a corpus of home maintenance instructions (RD), and one month of queries submitted to The Electric Monk, a natural language search system. Bierner quotes the following precision scores: BNC – 47%, RD – 57.8%, Monk – 78.3%. The recall score is the same, because the procedure identified an anchor for all instances of "other".

These results are promising, but there is plenty of room for improvement. First, the algorithm concentrates on cases with explicitly given nominal antecedents and ignores adjectival, propositional, and utterance antecedents, which account for a total of 6.5% antecedents in the study corpus. Furthermore, the algorithm does not intend to handle anchors that require inference (beyond a default heuristic for situational anchors), e.g., discourse topic (3%), inferred anchors (6%) and those involving metonymy and metaphor (no figures from the study corpus are available at present).

Second, empirical findings reported in Sections 2.3 and 2.4 – the low frequency of pronominal antecedents and the ability of "other" to refer to an entity mentioned as far as three and more sentences away – suggest that "other" is able to access discourse entities that may no longer be in the center of attention. Such antecedents will not be found by Bierner's algorithm, which relies on the idea of discourse salience. This is perhaps one of the explanations of why the algorithm's precision scores drop with the increased complexity of texts. To estimate the impact of salience on the resolution of "other NPs", I am currently running an experiment on the GNOME corpus which has been marked for various features that have been noted as relevant to determine salience. (For further information about the GNOME project see http://www.iccs.informatics.ed.ac.uk/~poesio/ under Projects.)

The third comment concerns the efficiency of Bierner's search algorithm. Sentences in the BNC texts turn out to be quite long, with many noun phrases. The algorithm considers each one in turn, in the order in which they are supplied by a salience module, checking whether they are consistent with the properties of the anaphor. This seems somewhat inefficient. Instead, I suggest to use lexical information available from the "other NP" to *guide* the search. The idea is the following: if the phrase under consideration is "other Xs", then only its hyponyms would qualify as candidates; other NPs do not need to be considered. Cases that involve redescription (e.g., "dogs and other pets" rather than "dogs and other canines") are more complex and require further empirical work. An experiments is under way to examine search complexity for such cases.

Lexical approaches to textual cohesion and reference resolution, similar to the one outlined here, have been implemented and tested by e.g., (Morris and Hirst, 1991) and (Harabagiu, 1998). Harabagiu reports that her approach, when combined with attentional constraints, proved to render a better performance for two classes of anaphora – pronouns and definite descriptions.

4 Summary

This paper presented a first-pass corpus analysis of comparative "other NPs" from the British National Corpus. The study corpus was annotated with easily observable features such as syntactic type of the antecedent, its proximity to the "other NP" (measured in sentence units), and type of lexico-semantic relation between "other" and its anchor.

The data analysis revealed several statistically significant patterns (and more might emerge in future work), which will be used to design a resolution algorithm that complements the existing approach of (Bierner, 2000). (Algorithm implementation and evaluation are subjects for future work.)

The major difference between the approach outlined in this paper and Bierner's is that my algorithm does not rely on the notion of discourse salience. Two observations in the study data

seem to suggest that "other" is able to access discourse entities that are no longer in the center of attention of the speaker and hearer. The first one is the low frequency of pronominal antecedents of "other" (6%). The second is that 10% of definite NP and 9% of proper name antecedents – two of the three most frequent antecedent types – are found as far as three and more sentences away from "other".

This study has also shown that anchors of "other" can be evoked by a larger spectrum of expressions than previously noticed. For instance, clauses, utterances, discourse segments, and strings of text, as well as the utterance situation, can render discourse anchors of "other". Furthermore, both explicitly evoked referents and those that are mediated by the text or utterance situation might involve a variety of inferential processes, such as e.g., metonymy and metaphor. Practical applications, e.g., Information Extraction and Automatic Summarisation, require resolution algorithms that can handle all types of discourse anchors. The statistical results derived from this study is a first step towards such algorithms.

Acknowledgements

I am grateful to Bonnie Webber, Ivana Kruijff-Korbayová, Katja Markert, Elisabet Engdahl, and the anonymous reviewer for comments and fruitful discussions of this paper. Many thanks are also due to audiences in Edinburgh, Göteborg, Toronto, Stockholm, and Umeå, where the first drafts of this paper were presented.

References

Bierner, G. (2000). *Alternative phrases: Theoretical analysis and practical applications.* PhD thesis, Division of Informatics, University of Edinburgh.

Bierner, G. (2001). Alternative phrases and natural language information retrieval. In *Proceedings of ACL-EACL 2001*, Toulouse, France.

Byron, D. K. and Allen, J. F. (1998). Resolving demonstrative anaphora in the TRAINS93 corpus. In *Proceedings of DAARRC2 - Discourse, Anaphora and Reference Resolution Colloquium.* Lancaster University.

Carter, D. (1987). *Interpreting anaphors in natural language texts.* Ellis Horwood Limited, John Wiley and Sons, Chichester.

DeCristofaro, J., Strube, M., and McCoy, K. F. (1999). Building a tool for annotating reference in discourse. In *Proceedings of the ACL99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 54–62.

Fraurud, K. (1992). *Processing noun phrases in natural discourse.* PhD thesis, Stockholm University.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherehce of discourse. *Computational Linguistics*, 21(2):203–225.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English.* Longman, London.

Harabagiu, S. M. (1998). *WordNet-based inference of textual context, cohesion and coherence.* PhD thesis, University of Southern California.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44:311–338.

Keller, F., Corley, M., Corley, S., Crocker, M., and Trewin, S. (1999). Gsearch: A tool for syntactic investigation of unparsed corpora. In Uszkoreit, H., Brants, T., and Krenn, B., editors, *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, pages 56–63, Bergen.

Miller, G. A. (1998). Nouns in WordNet. In Fellbaum, C., editor, *WordNet: An Electronic Lexical database*, pages 23–46, Cambridge, Mass. MIT Press.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comupational Linguistics*, 17(1):21–48.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, 1:75–116.

Steedman, M. (1996). *Surface structure and interpretation.* Number 30 in Linguistic Inquiry Monograph. MIT Press, Cambridge, Mass.

Steedman, M. (2000). *The syntactic process.* MIT Press, Cambridge, Mass.

Strube, M. (1998). Never look back: An alternative to centering. In *Coling-ACL'98: Proceedings of the 17th Int. Conference on Computational Linguistics and the 36th Annual Meeting of the ACL*, pages 1251–1257.

van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4).