Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies

Ruslan Mitkov*, Richard Evans*, Constantin Orasan*, Catalina Barbu*, Lisa Jones*, and Violeta Sotirova†

* School of Humanities, Languages and Social Sciences
University of Wolverhampton
Wolverhampton WV1 1SB
United Kingdom
{R.Mitkov, in6087, in6093, in6465, in6942}@wlv.ac.uk
†University of Manchester
vsotirova@hotmail.com

Abstract

The paper summarises the work of the Research Group in Computational Linguistics at the University of Wolverhampton towards the production of much needed annotated resources for evaluation and training of anaphora resolution systems. In particular, it describes the annotating tools developed to support the annotation, the corpora annotated and the annotation strategy adopted. Finally, future plans are outlined.

1. The need for annotated corpora in anaphora and coreference resolution

Since the early 1990s, research and development in both anaphora and coreference resolution has been benefiting from the availability of corpora, both raw and annotated. Raw corpora are commonly available and have been found very useful in a wide range of NLP tasks, including lexicography, term extraction, and clustering methods. However, they have so far made only a limited contribution to the process of anaphora resolution, (Dagan and Itai, 1990; 1991) using them in order to extract collocation patterns for use by their system.

Corpora annotated with anaphoric or coreferential links are not widely available, even though they are much needed for different methods in anaphora/coreference resolution systems. Corpora of this kind have been used in the training of machine learning algorithms (Aone and Bennett, 1995) or statistical approaches (Ge et al., 1998) to coreference resolution. In other cases, they were used for optimization of existing approaches (Orasan et al., 2000) and their evaluation (Mitkov et al., 1999). The automatic training and evaluation of anaphora resolution approaches require that the annotation covers anaphoric or coreferential chains and not just single anaphor-antecedent pairs, since the resolution of a specific anaphor would be considered successful if any preceding non-pronominal element of the anaphoric chain associated with that anaphor, is identified. Unfortunately, anaphorically or coreferentially annotated corpora are not widely available, and those that exist are not of a large size. In 1999, the Research Group in Computational Linguistics at the University of Wolverhampton embarked upon an initially small-scale, but steadily expanding project aiming to partially satisfy this need.

2. Basic notions: anaphora and coreference

We define *anaphora* as the linguistic phenomenon of pointing back to a previously mentioned item in the text as opposed to *coreference*, which is the act of referring to the same referent in the real world. Note that not all varieties of anaphora have a referring function, such as verb anaphora, for example.

Tony McEnery organised$_i$ DAARC-3 as he did$_i$ DAARC-1.

Also, the anaphor and the antecedent may refer but may still not be coreferential as in the case of identity-of sense anaphora.[1]

The man$_i$ who gave his$_i$ paycheck$_j$ to his$_i$ wife was wiser than the man$_k$ who gave it$_j$ to his$_k$ mistress.

as opposed to identity-of-reference anaphora

This man$_i$ gave his$_i$ paycheck to his$_i$ wife.

Bound anaphora is another example where the anaphor and the antecedent are not coreferential,

Every participant$_i$ had to present his$_j$ paper.

Anaphora normally operates within a document (e.g. article, chapter, book), whereas coreference can be taken to work across documents. We have seen that there are varieties of anaphora that do not involve coreference; it is also possible to have coreferential items that are not anaphoric with *cross-document coreference* being an obvious example: two mentions of the same person in two different documents will be coreferential, but will not stand in an anaphoric relation.

Nominal anaphora arises when a referring expression - pronoun, definite noun phrase, or proper name, has a non-pronominal noun phrase as antecedent[2]. Broadly speaking, there are considered to be two types of nominal anaphora:

[1] In identity-of-sense anaphora, the anaphor and the antecedent do not correspond to the same referent in the real world but to ones of a similar description

[2] There are other types of anaphora in which anaphors point back to non-nominal words or phrases such as verbs or clauses. In this work, we do not treat this class

direct and *indirect*. *Direct anaphora* links anaphors and antecedents by such relations as identity, synonymy and specialisation. In contrast, *indirect anaphora* links anaphors and antecedents by relations such as meronymy/holonymy. Resolution of indirect anaphora normally requires the use of domain or world knowledge. Indirect anaphora is also known as *associative* or *bridging* anaphora. For more on the notions of anaphora and coreference and on the different varieties of anaphora see (Hirst, 1981; Mitkov, 2001).

Our project addresses the most crucial type of anaphora to NLP applications – that of identity-of-reference direct nominal anaphora, which can be regarded as the class of single-document identity coreference. This most frequently occurring class of anaphora has been researched and covered most extensively, and is the best understood within the field.

3. The complexity of the task

Annotating anaphora is a notoriously difficult, time-consuming and labour-intensive task even when focusing on one single variety of the phenomenon. Consider the case of demonstrative anaphora - it is well known that when the antecedent is a text segment longer than a sentence, it is often difficult to decide exactly which text portion represents the antecedent.

Also, a number of studies (Hirschman, 1997; van Deemter and Kibble, 1999) demonstrate that even when limited to identity coreference, the decisions as to what to annotate and how to annotate are not simple. As a consequence, the annotation process is often considered to be far from reliable in that interannotator agreement may be disappointingly low.

(van Deemter and Kibble, 1999) show that in addition to the numerous blurred cases which represent a formidable challenge to annotators, the coreference annotation task suffers from terminological confusion. In particular, they criticise the MUC coreference annotating scheme which they claim 'goes beyond annotation of coreference as it is commonly understood' since it marks non-referring NPs (which therefore cannot co-refer) such as quantifying NPs (e.g. *every man, most computational linguists*) as part of the coreferential chains.[3] The authors also express reservation regarding the marking of indefinite NPs and predicate NPs as possibly coreferential arguing that if in the example

Henry Higgins, who was formerly sales director of Sudsy Soaps, became president of Dreamy Detergents

Henry Higgins, sales director of Sudsy Soaps and *president of Dreamy Detergents* are all marked as standing in the IDENT relation, and if two NPs should be recorded as coreferential if the text asserts them to be coreferential at any time[4], then one could conclude that *Henry Higgins* is presently *sales director of Sudsy Soaps* as well as *president of Dreamy Detergents* which is not what the text asserts.[5]

In answer to such criticisms, it should be emphasised that the MUC-7 coreference annotation scheme was designed in order to produce resources for practical automatic evaluation of coreference resolution systems, rather than to capture the full range of coreferential relations between elements. As we discovered and discuss further in Section 5, even when following a scheme that provides coverage of only a limited range of textual relations, such as the MUC-7 one, the coreferential annotation of corpora is still an incredibly slow process requiring a huge amount of concentration.

4. Overview of existing annotated resources, annotation schemes and annotating tools

4.1. Corpora annotated with anaphoric or coreferential links

One of the few anaphorically annotated resources, the *Lancaster Anaphoric treebank* is a 100 000 word sample of the Associated Press (AP) corpus (Leech and Garside, 1991), annotated with the UCREL anaphora annotation scheme and featuring a wide variety of phenomena ranging from pronominal and NP anaphora to ellipsis and the generic use of pronouns.

The *annotated data produced for the MUC coreference task* amounts to approximately 65 000 words[6] and lists coreferential chains from newswire reports on subjects such as corporate buyouts, management takeovers, airline business and plane crashes.[7] (Popescu, 1998) reports *two texts annotated for coreferential links in French:* the first one, marked up in both MUC's and Bruneseaux and Romary's schemes (see Section 4.2), is part of a short story by Stendahl (total of 638 referential expressions), whereas the second is part of a novel by Balzac and follows Bruneseaux and Romary's scheme (total of 3812 referential expressions). Annotated data for French are being produced within a current project conducted jointly by the *University of Stendahl, Grenoble and Xerox Research Centre Europe* (Tutin et al., 2000) which is expected to deliver a 1 million word corpus annotated for anaphoric and cataphoric links. The annotation is limited to anaphor-closest antecedent pairs rather than full anaphoric chains[8] and involves 3rd person personal pronouns, possessive pronouns, demonstrative pronouns, indefinite pronouns, adverbial anaphors, zero noun anaphors but not definite noun phrases. Finally, as a consequence of the increasing number of projects in multilingual anaphora resolution, the need for parallel bilingual and multilingual corpora annotated for coreferential or anaphoric links has become obvious. To the best of our knowledge there are no such corpora available yet apart from a small-size *English-Romanian*

[3] The authors argue that MUC mixes up coreferential and anaphoric relations

[4] According to the MUC-7 guidelines

[5] van Deemter and Kibble propose alternative solutions in their paper

[6] This figure is based on data/information kindly provided to us by Nancy Chinchor

[7] Some of the articles are also about reports on scientific subjects. Management of defence contracts is covered and there are also reports on music concerts, legal matters (lawsuits, etc.) and broadcasting business

[8] This limitation makes the corpus more suitable for theoretical linguistic research than for evaluation and testing anaphora resolution system where full anaphoric or coreferential chains are needed

corpus developed for testing *a bilingual coreference resolution system* (Harabagiu and Maiorano, 2000) and a small *English-French bilingual corpus* developed at *the University of Wolverhampton* (see section 5.3.3).

4.2. Annotation schemes

In recent years, a number of corpus annotation schemes for marking up anaphora have come into existence. The first and one of the most sophisticated schemes is the *UCREL scheme* initially developed by Geoffrey Leech (Lancaster University) and Ezra Black (IBM). It allows the marking of different varieties of anaphora including ellipsis, but also non-referential use pronouns. In addition, annotator uncertainty can be marked. Special symbols added to anaphors and antecedents can encode the direction of reference (i.e. anaphoric or cataphoric), the type of cohesive relationship involved, the antecedent of an anaphor, as well as various semantic features of anaphors and antecedents. One drawback of the UCREL scheme is that it is difficult for computers to process texts marked using this annotation.

The SGML-based *MUC annotation scheme* (Hirschman, 1997) which is seemingly the least ambitious one in terms of coverage, has been used by a number of researchers to annotate coreferential links (Gaizauskas and Humphreys, 1996; 2000; Mitkov, Orasan, and Evans, 1999). In this scheme, the attribute ID uniquely denotes each string in a coreference relation, REF identifies which string is coreferential with the one which it tags, TYPE - indicates the type of relationship between anaphor and antecedent and the TYPE value 'IDENT' indicates the identity relationship between anaphor and antecedent. The MUC scheme only covers the identity (IDENT) relation for noun phrases and does not include other kinds of relations such as part-of or set membership. In addition to these attributes, the annotator can add two more, the first of which is MIN, which is used in the automatic evaluation of coreference resolution systems. The value of MIN represents the smallest continuous substring of the element that must be identified by a system in order to consider a resolution correct. Secondly, the attribute *STATUS* can be used and set to the value 'OPT'. This information is used to express the fact that mark-up of the tagged element is optional.

Dates, currency expressions and percentages are considered noun phrases. The MUC scheme also stipulates which noun phrases should be marked up as coreferential and when, but has been criticised for some inconsistencies on that matter. In spite of its imperfections, the MUC scheme has the strength of offering a standard format. Also, although it has been designed to mark only a small subset of anaphoric and coreferential relations, the SGML framework does provide a useful starting point for standardisation of different anaphoric annotation schemes.

The *MATE scheme for annotating coreference in dialogues* (Davies et al., 1998) draws on the MUC coreference scheme, adding mechanisms for marking-up further types of information about anaphoric relations as done in the UCREL, DRAMA and Bruneseaux and Romary's schemes. In particular, this scheme allows for the mark up of anaphoric constructs typical in Romance languages such as clitics and of some typical dialogue phenomena. The scheme also provides for the mark up of ambiguities and misunderstandings in dialogue.

The XML-based scheme proposed in (Tutin et al., 2000) supports the annotation of a variety of anaphoric relations such as coreference, set membership, substitution, sentential anaphora and indefinite relations which includes all cases not covered by the first four types such as bound anaphora.[9]

Other well known schemes include *de Rocha's* (1997) *scheme for annotating spoken Portuguese*[10], *Botley's* (1999) *scheme for demonstrative pronouns*, *Bruneseaux and Romary's scheme* (1997), *the DRAMA scheme* (Passonneau and Litman, 1997) and *the annotation scheme for marking up definite noun phrases* proposed by (Poesio and Vieira, 1998; 1999).

4.3. Annotating tools

In order to help the human annotator it is necessary to provide him/her with a tool which makes it possible to quickly identify the entities in the discourse and the relations between them. A good graphical interface offers the human annotator trouble-free and efficient interaction with the annotated text and the tool must be easy to use; thus minimizing the time required to learn how to use it. It should also display the resulting annotation in a way that is easy for a user to interpret, hiding unnecessary or hard-to-read aspects of the annotation, such as raw SGML encoding, from the user.

The first tool for annotation of anaphoric links, *XANADU*, written by Roger Garside at Lancaster University, is an X-windows interactive editor that offers the user an easy-to-navigate environment for manually marking pairs of anaphors-antecedents within the *UCREL* scheme (Fligelstone, 1992). In particular, *XANADU* allows the user to move around a block of text, displaying around 20 lines at a time. The user can use a mouse to mark any segment of text to which s/he wishes to add some labelling.

The *DTTool* (Discourse Tagging Tool), (Aone and Bennett, 1994), enables the annotation of anaphoric relations in Japanese, Spanish and English. This is done in a graphical manner - by the colour coding of different types of anaphors (e.g. 3rd person pronoun, definite NP, proper name, zero pronoun, etc.) and antecedents which are displayed on the screen with arrows linking them. The annotated data can be viewed in five different modes: all tags, each anaphor-antecedent pair, all anaphor-antecedent pairs of the same type, all anaphoric chains and the text without any tags.

The *Alembic Workbench*, (Day et al., 1997), was developed at MITRE and has been used among other things, to mark-up coreference relations. In the coreference annotation task the workbench features a window that produces a sorted list of all tagged elements to facilitate the finding of coreferring expressions. The semi-automatic mode extends to simple annotating tasks such as tagging named entities. The Alembic Workbench offers a choice of tag sets, including all those necessary

[9] The original implementation of the scheme does not include anaphoric definite noun phrases

[10] One particularity of this scheme is that it allows the annotator to encode the kind of information that is necessary to resolve the anaphoric expressions

for the MUC scheme and provides a graphical interface which allows the modification of existing tags and the addition of new ones. Users of the system are also able to construct their own task-specific annotation schemes.

Referee is a discourse annotation and visualisation tool that operates in three modes – reference mode, segment mode and dialogue mode (DeCristofaro et al., 1999). In reference mode, the user can mark words or expressions by associating features (e.g. syntactic function, distance, pronominalisation, definiteness etc.) with each of them and assigning coreference. In segment mode, the user can partition the text into arbitrarily nested and overlapping segments, whereas the dialogue mode enables him/her to code a dialogue by breaking it into turns.

4.4. CLinkA

At the time that our project for building the annotated corpus started, the only tool that was available for facilitating the annotation task was the Alembic Workbench. Bugs in the version available at that time and a discontinuation of support for the Windows version motivated our decision to develop our own tool for coreference annotation.

The *Coreference Links Annotator*, referred to as *CLinkA*, (Orasan, 2000), is a platform and language independent tool developed by our group in order to aid the task of coreference resolution. To ensure *ClinkA*'s language independence, we decided to implement it in Java, due to its support for Unicode. Moreover, being developed in Java, the tool is platform independent, and may be run on any system that has a Java machine. The tool was tested on three languages which use Latin characters: English, Spanish and Romanian and one which uses Cyrillic characters: Bulgarian.

Given that when we started the annotation project we urgently needed an operational tool, *CLinkA* in its initial version, only supports the MUC-7 annotation scheme. In a future release we intend to extend the program, by allowing the user to define his/her own annotation schemes.

Aiding the annotation process, the tool offers a user-friendly graphical interface, in which different types of entities are marked using different colours. In this way, the user can quickly identify what s/he is looking for. The process of annotation is kept as simple as possible in order to make the program useful also for people who are not computer literate. In order to mark an entity, its boundaries have to be identified using the mouse and in cases when it has to be added to a chain, that chain has to be indicated by clicking on an element already in the chain. We noticed, however, that even with this tool, the process of annotation is quite time consuming, therefore we added a mode for semi-automatic tagging. In this mode, once the user has assigned an entity to a chain, all identical textual strings are proposed to be included in that chain. For entities that contain more than one word, this usually works with high accuracy, but it does not deliver acceptable results for single word entities. In all cases, the user decides if the program's suggestion is correct or not. Further ways of speeding up the process are discussed in section 7.2.

5. Developing annotated data in Wolverhampton

5.1. The sensible trade-off

Given the complexity of the anaphora and coreference annotation task, we have decided to adopt a less ambitious but clearer approach than some of those presented in Section 4.2 as to what variety of anaphora to annotate. This move is motivated by the fact that (i) annotating anaphora and coreference in general is a very difficult task and (ii) our aim is to produce annotated data for the most widespread type of anaphora which is the main focus in NLP: that of identity-of-reference direct nominal anaphora featuring a relation of coreference between the anaphors (pronouns, definite descriptions or proper names) and any of their antecedents (non-pronominal NPs).[11] We annotate identity-of-reference direct nominal anaphora, which includes relationships such as specialisation, generalisation and synonymy, but excludes part-of and set membership relations that are considered instances of indirect anaphora. Whilst we are aware that such a corpus will be of less interest in linguistic studies, we believe that the vast majority of NLP work on anaphora and coreference resolution (and all those tasks which rely on it) will be able to benefit from this corpus by using it for evaluation and training purposes. Therefore, we believe that the trade-off of a wide coverage, but complicated and potentially error-prone annotation task with low-consistency across annotations for a simpler, but more reliable annotation task with a NLP-orientated end product is a worthwhile endeavour.

5.2. Adopted scheme: MUC-7

For our annotation, we adopted the MUC–7 annotation scheme for coreference (Hirschman, 1997). Various researchers have commented on the disadvantages of that scheme, pointing to the shortcomings inherent in the restriction of the coreference task to the relation of identity only. In addition, the MUC-7 annotation scheme assumes markable elements to be continuous strings. It is therefore impossible to capture relations between plural pronouns and discontinuous antecedents as in

John$_{i}$ *goes to college on Mondays,* *Mary*$_{j}$ *goes on Tuesdays, and they*$_{ij}$ *both go on Wednesdays.*

Here, a relation holds between *they* and the conjunction of *John* and *Mary* but *John* and *Mary* are not coreferential. Nevertheless, adoption of the MUC-7 scheme is practical because of the relatively high level of development of its annotation guidelines and the fact that it is coded in SGML, which facilitates subsequent processing of annotated documents. In addition, as mentioned in Sections 4.3 and 4.4, there are currently a number of tools available to facilitate more rapid annotation using the MUC-7 scheme.

[11] Since the task of anaphora resolution is considered successful if any element of the anaphoric (coreferential) chain preceding the anaphor is identified, our project addresses the annotation of whole anaphoric (coreferential) chains and not only *anaphor-closest antecedent* pairs

TEXT	GENRE	WORDS	ELEMENTS	CHAINS	ANNOTATION (NOMINAL IDENTITY-OF-REFERENCE DIRECT ANAPHORA)
Aiwa	VCR technical manual	6723	1630	914	Full
FreeBSD	Software technical manual (extract)	1999	473	270	Full
Hinari	TV technical manual	2821	673	381	Full
Macint	Computer technical manual	15131	233	65	Pronouns - Antecedents only
Panason	TV technical manual	4843	1265	726	Full
Urbancorpus	Home improvements safety manual (extract)	2215	453	372	Full
Winhelpfile	Software technical manual (extract)	2882	673	463	Full

Table 1: Characteristics of some of the existing corpora

5.3. Brief description of the corpora

At Wolverhampton, the scheme described in Section 5.2 was instituted for the annotation of our corpora. Some of the annotated texts are presented in Table 1 with information as to their size, genre, and information about the numbers of markable elements that they contain.

It should be noted that these resources have certain limitations. Firstly, some of the texts were only annotated by one person and no inter-annotator validation was performed for them. Secondly, in the initial stages, they were not annotated with the assistance of guidelines; instead, an intuitive annotation process was performed. In light of this, it is expected that inter-annotator agreement in respect of these corpora will be at a low level. The shortcomings of this annotation have been addressed by the proposal and adoption of detailed annotation guidelines as described in Section 6.1.

5.3.1. Annotating full coreferential chains

We are concerned with identifying the full coreferential chains in the corpora. This task depends on identifying the markables that represent initial mentions of entities in the texts and then identifying the markables that are coreferential with them. As in the MUC-7 scheme, we assume that the coreference relation is transitive and each element in a coreference chain is marked as identical to the initial mention. Under this approach to annotation, the annotator must explicitly identify every single markable in the text. The resulting corpora are most suitable for evaluation of all systems that treat identity-of-reference direct nominal anaphora, regardless of the actual algorithms on which those systems are based.

Full annotation of coreferential chains is a time-consuming process. In an experiment over one text, we found that annotators were able to assign 288 elements to 220 chains[12], covering only 2051 words per hour. Reference to annotation guidelines improves the granularity and accuracy of the annotation process. Providing annotators with reference to guidelines, we found on average that they assigned 315 elements to 250 chains, covering 1411 words per hour.

[12] Here, the number includes partial chains and single element chains

5.3.2. Fast-track annotation

As mentioned previously, the annotation of full coreferential chains is a time consuming task. In some cases, like pronominal anaphora resolution, only parts of these chains are useful. Therefore, in light of the time constraints imposed by research projects it can be beneficial to seek faster methods for producing the necessary resources.

We primarily wanted to use the annotated corpora in the processes of evaluation and optimisation of our approach to pronominal anaphora resolution, (Orasan et al., 2000), but such processes require a large amount of annotated material. We thus found it suitable to produce a relatively large amount of annotated data by adopting a much less labour-intensive annotation method.

To this end, we developed a system that first uses Conexor's FDG Parser (Tapanainen and Jarvinen, 1997) to lemmatize the texts and extract pronouns from them. In addition, the extracted pronouns are associated with their NP candidates, together with the positions of these elements. The resulting file is a list of pronouns, each pronoun followed by a list of candidates taken from the preceding text of the same paragraph as the pronoun[13]. Each candidate is automatically printed next to a '-' symbol that indicates that it is not the antecedent of the pronoun. The task of the annotator is simply to replace '-' with '+' for candidates that are the correct antecedents. In cases of ambiguity, the '?' symbol is used instead. The system facilitates the annotation task by also printing the lemmatised paragraph in which each pronoun appears.

Adopting this strategy, the annotation process is changed to the identification of local parts of chains that contain anaphoric pronominal elements. Fast-track annotation requires the classification of NPs as antecedents or non-antecedents of pronouns. It was found that an average of 113 pronouns, 944 candidates, and 148 antecedents; covering 10862 words; could be marked per hour[14].

[13] Our assumption is that the list of candidates selected by the algorithm will always be a subset of the annotated list, regardless of any changes in the resolution algorithm

[14] It should be noted that in some cases, the same candidate will have to be annotated more than once because it may occur in the candidate sets of several pronouns

DO:	DO NOT:
(i) annotate identity-of-reference direct nominal anaphora	(i) annotate indefinite predicate nominals that are linked to other elements by perception verbs as coreferential with those elements
(ii) annotate definite descriptions which stand in any of the identity, synonymy, generalisation, specialisation, or copula relationships with an antecedent	(ii) annotate identity-of-sense anaphora
(iii) annotate definite NPs in a copula relation as coreferential	(iii) annotate indirect anaphora between markables
(iv) annotate appositional and bracketed phrases as coreferential with the NP of which they are a part	(iv) annotate cross-document coreference
(v) annotate NPs at all levels from base to complex and co-ordinated	(v) annotate indefinite NPs in copula relations with other NPs as coreferential
(vi) familiarise yourself with the use of unfamiliar, highly specialised terminology by search through the text	(vi) annotate non-permanent or "potential" coreference between markables
	(vii) annotate bound anaphors
	(viii) consider gerunds of any kind markable
	(ix) annotate anaphora over disjoined antecedents
	(x) consider first or second person pronouns markable

Table 2: Annotation Guidelines

The disadvantage is that the coreferential chains that can be inferred from such data are highly unlikely to be complete. While useful in the evaluation of our pronominal anaphora resolution system, it is not possible to apply such resources to other systems, in general.

5.3.3. Parallel corpora

One of our research interests is multilingual anaphora resolution. We are currently focusing on English and French and for supporting the anaphora resolution system we are producing a parallel bilingual corpus annotated with coreferential links. The texts that we are working on are extracted from the Canadian corpus BAF, containing scientific, technical and narrative texts. The main advantage of these texts is that they are already aligned at sentence level and they are also rich in pronouns. The corpus annotated so far is quite small, consisting of about 10000 words with 150 pronouns in each language.

6. Annotation strategies

6.1. Brief outline of the annotation guidelines

6.1.1. The need for revised annotation guidelines

Our annotation guidelines are based on those proposed by (Hirschman, 1997) for the MUC-7 coreference task. The differences between our guidelines and the MUC-7 ones can be divided into two main types.

Firstly, although the MUC-7 documentation provides guidance for a variety of relations between elements with respect to coreference, we found the treatment of some of the interesting relations objectionable. To illustrate, we do not consider indefinite appositional phrases coreferential with the phrases that contain them. For the purpose of identity-of reference direct nominal anaphora resolution, we do not consider the relation between these elements to be definite enough to encode. We decided not to consider gerunds of any kind to be markable. The primary motivation for this is that we intend to keep the annotation process as simple as possible and we consider that requiring annotators to only identify particular types of gerund as markable adds unnecessary scope for inter-annotator disagreement to the task.

Secondly, we observed a number of relations and phenomena in our corpora for which no guidance had been formulated in the MUC-7 guidelines. For example, structures like [V [NP_1] as [NP_2]] as in

[*elect* [*John Prescott*] *as* [*Prime Minister*]], or

[*use a* [*diagonal linear gradient*] *as* [*the map*]]

should not be marked coreferential. Another case concerned the annotation of conjoined elements in which different words are missing. We decided that constructs such as

…the pixels' luminance or saturation is important… The pixels' saturation must…

should be marked so as to convey the structure

…[[*the pixels' luminance*]$_i$ *or* [∅ ∅ *saturation*]$_j$]$_k$ *is important…* [*The pixels' saturation*]$_j$

with coreference holding between both the second phrase in the conjunction and the other NP. We consider these constructs to demonstrate NP conjunction with ellipsis before the second noun head. This contrasts with those cases in which the conjoined phrase has only one explicit noun head, as in

…if [[*an NTSC* ∅]$_i$ *or* [*PAL monitor*]$_j$]$_k$ *is being used …* [*The NTSC monitor*]$_m$…

in which the first conjoined phrase lacks a noun head and we do not consider coreference to hold between it and the subsequent complete noun phrase. Exclusively disjoined pronouns such as *he/she* are marked as single units, given that their antecedents in the text do not usually carry

SAMPLE	#WORDS	GUIDELINES	AVERAGE #CHAINS	AVERAGE #ELEMENTS	AVERAGE PROPORTION SHARED ELEMENTS	AVERAGE F-MEASURE (%)	#ANNOTATORS
1	3374	N	163	215.67	0.66	65.25	3
2	2117	Y	374	471.5	0.72	62.16	2
3	1315	Y (2 pass: initial mentions first)	164	248	0.67	51.26	2

Table 3: Files annotated with CLinkA

specific gender feature information (e.g. *the user of the machine* or *a teacher of English*).

In addition, as the texts of the corpora are derived from web documents, they are often found to contain extraneous material such as contact addresses and menus. Our guidelines instruct annotators to disregard such incoherent occurrences of apparent markables.

We wanted to ensure that our guidelines were practical to consult and we therefore restricted the amount of information that they contain to crucial guidance on the most highly ambiguous and frequently occurring annotation cases. At the present time the guidelines span just three pages, whereas the MUC-7 guidelines span seventeen. The guidelines are summarised in Table 2.

As mentioned earlier, guidance is provided for a number of other difficult cases including the identification of markable elements in web page documents, the treatment of conjoined NPs with head or modifier ellipsis, and the treatment of exclusively disjoined pronouns with differing gender feature information.

6.2. Quality checks

In assessing the validity of our annotations, we require that each text be independently annotated by at least two annotators. Comparison of these annotations is facilitated by a program that matches all annotations and flags up instances marked up differently by the annotators. The program works by extracting the full coreference chains from two annotated files and then producing the chains that are present in one file but are not identical to any chains in the file being compared. Similarly, differing elements are written and the number of elements shared between the files is returned. This allows a qualitative assessment of the differences between the annotations as well as subsequent discussion and adjudication. (Hirschman et al., 1998) compared annotations using the measures of precision and recall. There, one annotator is defined as a *key* and the other as a *response*. The response is then assessed with respect to the key. Precision is defined as the ratio of the number of coreference links in the intersection of the key and the response to the number of links in the response. Recall is the ratio of the number of coreference links in the intersection of key and response to the total number of links in the key. We implemented a program that produces all the links that can be inferred from the coreference chains in annotated texts. The set of links associated with a file is then compared with the set produced from the text that is taken as the key.

6.3. Interannotator agreement

A minimum of two annotators marked some new samples of a text from the domain of software instruction manuals. In one case, three annotators marked a sample. For each pair of annotators, we then obtained statistics crucial for measuring interannotator agreement. These were the proportion of elements marked by both annotators, and the measures necessary in order to evaluate the coreferential links identified by one annotator as a response against the links identified by another annotator as the key (Hirschman et al., 1998). The scores used to assess interannotator agreement over the different extracts are presented in Table 3. Computing the proportion of shared elements, the following formula was used:

$$\mu = \frac{2*C}{A_1 + A_2}$$

where A_1 and A_2 are the number of entities in the first and second annotations, respectively, and C represents the number of entities common to both annotations.

In computing F-measure, we set $\alpha = 0.5$, giving equal importance to precision and recall. Due to the fact that precision, recall, F-measure, and μ can only capture agreement between a pair of annotators, we present the average (mean) values for the two pairs involved in annotation of sample 1. We note that adoption of the guidelines presented in Section 6.1 over samples 2 and 3 gave no significant improvement in terms of inter-annotator agreement although it did improve the granularity of the annotation. For a different text, this time strictly adopting the guidelines, the F-measure was at the level of 62.16%. Combining the guidelines with a two-pass annotation strategy; where in the first pass, initial mentions are marked and in the second pass, coreferential mentions are assigned to the appropriate chains; also did not improve interannotator agreement. Unfortunately, there is no automatic way to perform a qualitative assessment of the resulting annotations.

6.4. How agreement can be improved

The data on discrepancies between chains and elements derived from two annotations of the same text, resulting from the program described in Section 6.2, was consulted and used as the basis for more effective

annotation strategies and treatments of problematic features of the texts.

In the following, it is important to note that most annotators find the annotation task to be tedious and also to require high levels of concentration. Lapses in concentration during prolonged annotation sessions can easily lead to the omission of elements and most inter-annotator discrepancies fall into this class.

Qualitatively, most of the discrepancies between annotations that occurred once the guidelines had been adopted were broadly due to two reasons.

Firstly was the perceived phenomenon of *unsteady references* in which a particular phrase such as *the image* is not felt by annotators to refer to the same entity throughout the entire course of the document. For example, certain actions on behalf of a textual agent may mean that one element, such as an electronic image, changes so much from section to section that these mentions cannot be considered coreferential. The discrepancies seem to arise when annotators differ with respect to the point in the text at which they no longer consider identical strings as coreferential. Resolution of this problem depends upon the ability of annotators to determine topic changes in the text in a more reliable and consistent way. Perhaps a preceding stage should be introduced into the annotation process in which annotators mark the perceived topic segments of a discourse and derive segment-internal coreference chains.

Secondly, discrepancies arose because of differences in the levels of specialised domain knowledge held by individual annotators. Given the technical nature of the corpora being annotated, we provided some guidance to help annotators deal with specialised terminology, suggesting a search through the texts by annotators in order to disambiguate the way in which previously unseen specialised terminology is used.

In general, we expect it to be useful to combine individual annotators' approaches to form a 'master' strategy for annotation, and the following tactics were suggested:

- Prior to annotation, print out the whole text to familiarise yourself with it, identifying all the noun phrases to be marked as either initial or subsequent mentions.

- Make a note of all troublesome or ambiguous cases and discuss them with other annotators to decide upon the best solution to tackle them in future.

- Ensure that the annotation is done in one intensive period, as sporadically annotating a file can lead to the annotator having to re-read the document for familiarisation several times.

In addition to these, individual annotators may find it beneficial to use additional tactics, such as a two-pass annotation process in which all initial mentions are identified in the first pass and all coreferential links are assigned in the second. However, we have found these to be best regarded as personal strategies, being of good value to some annotators, but not to others.

(Hirschman et al., 1998) noted that it was useful for annotators to undertake a two-pass annotation strategy, first identifying the positions of the heads of markable elements in a text and then assigning those elements to the appropriate coreferential chains. In our case, we do not encode the *MIN* feature, so annotators would need to mark full noun phrase elements in our annotation process. Such a strategy would be of limited use for us as we expect that the identification of complex elements will be as inconsistent across annotators regardless of the point at which it is performed. Marking the *MIN* feature and using it as the basis of evaluation has the benefit that noun heads are considerably easier to identify than complex noun phrases. We intend to investigate the importance of this advantage, weighing its benefits against the additional steps that are required to encode it during the annotation process.

7. Future work and ideas for improving the quality of the annotation

The task of developing coreferentially-annotated corpora is relatively recent and there are a large number of aspects of the process that can be improved. As a result of our project we envisage two ways of improving the usefulness of the corpora created. The first is to improve the annotation scheme used. The second way addresses certain difficulties in the annotation task and should lead to increased quality of the corpus produced.

7.1. Future work on the annotation schemes

Most of the criticisms of the MUC-7 scheme are due to its limited ability to capture the full range of phenomena that characterize the coreference relation. According to our preliminary evaluation the addition of new attributes and values to the existing tags will partially solve this problem. The MUC-7 scheme was initially developed as an extensible one. At present, the attribute *TYPE* accepts only one value, IDENT, but this can be changed to capture other relations between entities. The problem of discontinuous antecedents can be resolved by adding an additional relation that connects the discontinuous parts of the same entity. However, given that the main purpose of our corpora is automatic evaluation of systems, the requirement for these new features may not be so pressing.

(Bagga, 1998) makes a more practical proposal, where a classification of the coreferential links is made according to the resources necessary to resolve them. He points out that there are links that can be resolved easily using syntactic rules alone and others that require world knowledge to resolve. According to him, account should be taken of these classes when assessing the performance of systems. An extra attribute can be included in the tag that indicates the kind of information that is necessary in order to assign the element to a chain. This idea is not new in the field of annotation scheme design; (de Rocha, 1997) proposed a similar idea, but the advantage of Bagga's approach over de Rocha's is ease of processing.

By introducing these changes in the annotation scheme, the produced corpora will be useful not only for evaluation purposes, but also for theoretical research into anaphora and coreference. However, caution should be exercised in the introduction of new attributes, given the amount of work that is already required for annotation.

7.2. Further development of the coreference annotation tools

As we argued before, it is not possible to create coreferentially-annotated corpora without using tools specially designed for this task. Even when using these tools, the process of annotation takes a very long time. In this section we present possible ways of further developing the existing annotation tools in order to reduce the time and effort required for this process. The quality of the resulting corpora should improve as a direct result.

As we mentioned before, the existing tools provide a very simple supervised way of automatically adding elements to the existing chains in a text by assuming that all identical strings in a text have to be in the same coreferential chain. This is true for strings that contain several words, but totally wrong for many single-word units like "*it*" and other pronouns. An improved version of the rule takes into consideration the form of the entity that was annotated and does not make an automatic assignment of single-word units to a chain. It is obvious that further improvements can be proposed for this rule. As an example, by using a named entity recogniser, all identical company names, products, etc. that are one word in length can be considered members of the same chain.

Even with this improvement the time required for annotation does not decrease a great deal. According to our observations, most of the annotation time is spent individualising the markables in the text. Usually these markables are very well defined from a syntactic point of view. In the annotation scheme proposed in MUC-7, all noun phrases and pronouns are considered markables. It is straightforward to identify the pronouns, with reference to their surface form. At present, noun phrases, including the embedded ones, can be identified quite accurately by automatic means. Therefore a way to speed up the process of annotation is by automatically proposing the entities which have to be assigned to chains, and the task of the annotator will be to make sure that no entity is skipped and the boundaries of the currently identified entity are correct.

In order to help the annotator further, it is possible to identify with relatively high accuracy whether or not the identified entity is the initial mention of a chain, is non-coreferential, or is part of an existing chain. For identifying non-coreferential elements such as pleonastic *it*, machine learning (Evans, 2000) and rule-based (Paice and Husk, 1987) approaches were used. (Bean and Riloff, 1999) managed to identify non-anaphoric noun phrases with a high level of accuracy. These non-referential entities constitute either initial elements of chains, or isolated elements in the discourse. For an entity which is part of a chain, it is possible to propose to the annotator a list of the chains it is most likely to be a member of. This can be done using "shallow" pronoun and definite description resolvers.

For anaphora resolution, preference-based methods proved to be very easy to implement and fairly accurate. One of the existing methods could be incorporated into the system in order to propose the chains most likely to contain different elements. The main reason for proposing such a method, is the very low requirements of shallow anaphora resolution systems from a linguistic point of view and their robustness in the case of texts from an unseen genre (even if their performance is worse in those cases). As shown in (Orasan et al., 2000) these methods can be optimised on particular texts. One way to implement this change to existing annotation tools is by using intelligent background agents. One of the attributes of such an agent is that it can learn from examples and after it gains enough experience it starts manifesting. In our case, the agent makes adjustments to the scores used by the preference-based method on the basis of what the annotator marks. With the increase of entities marked, the accuracy of the agent in determining the most likely chains in which an element is in, increases. When this accuracy is high enough, it starts manifesting, proposing different chains with respect to elements to the user. Moreover, after annotating a file, the user can save the scores and use them later when annotating a different file from the same genre.

8. Conclusions

In this paper we presented some preliminary outcomes of our ongoing project in constructing coreferentially-annotated corpora. As we argued throughout the paper, this task is a very time-consuming one, requiring a lot of concentration. According to our experience the main difficulty comes from the way the annotation task is defined. We considered several annotation strategies as the basis for our approach and a new one, which represents a trade-off between approaches that attain high coverage and those that are clearly defined, was proposed. It has been designed more as a way of producing consistent resources for automatic evaluation and optimization of anaphora and coreference resolution systems, than as a way of capturing a wide range of linguistic phenomena. We carried out different experiments in order to determine which factors could improve the quality of the produced corpora. As future work, diverse ways for improving the usefulness and the quality of the resources were proposed.

9. References

Aone, C. and Bennett, S. W. 1994. Discourse tagging tool and discourse-tagged multilingual corpora. In *Proceedings of the International workshop on sharable natural language resources* (SNLR).

Bagga, A. 1998. Evaluation of Coreferences and Coreference Resolution Systems, In *Proceedings of the First International Conference on Language Resources and Evaluation*. ELRA. 563-566.

Bean, D. and Riloff, E. 1999. Corpus-Based Identification of Non-Anaphoric Noun Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Morgan Kaufmann.

Botley, S. 1999. *Corpora and discourse anaphora: using corpus evidence to test theoretical claims*. PhD thesis, University of Lancaster, UK.

Bruneseaux, F and Romary, L. 1997. Codage des références et coréférences dans les dialogues homme-machine. In *Proceedings of the Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing* (*ACH-ALLC '97*), 15-17. Ontario, Canada.

Dagan, I. and Itai, A. 1990. Automatic processing of large corpora for the resolution of anaphora references. In

Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1-3, Helsinki, Finland.

Dagan, I. and Itai, A. 1991. A statistical filter for resolving pronoun references. In Y.A. Feldman and A. Bruckstein (Eds) *Artificial Intelligence and Computer Vision*, 125-135. Elsevier Science Publishers B.V. (North-Holland).

Davies, S., Poesio, M., Bruneseaux, F., and Romary, L. 1998. *Annotating coreference in dialogues: proposal for a scheme for MATE*. First draft. Available at http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.html

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., and Vilain, M. 1997. Mixed-initiative development of language processing systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, (ANLP97)*, 153-164. Washington DC.

DeCristofaro, J., Strube, M., and McCoy, K. 1999. Building a tool for annotating reference in discourse. In *Proceedings of the ACL'99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, 54-62. College Park, Maryland, USA.

Evans, R. 2000. Applying Machine Learning Toward an Automatic Classification of *It*. To appear in *Literary and Linguistic Computing*. Oxford University Press.

Fligelstone, S. 1992. Developing a scheme for annotating text to show anaphoric relations. In *New Directions in English Language Corpora: Methodology, Results, Software Developments*. ed. by Leitner, G. 153-170. Berlin: Mouton de Gruyter.

Gaizauskas, R and Humphreys, K. 1996. Quantitative evaluation of coreference algorithms in an information extraction system. In Botley, S P and A M McEnery (eds.). 2000. *Corpus-Based and Computational Approaches to Discourse Anaphora*. Pages 143-167. Amsterdam: John Benjamins.

Ge, N., Hale, J., and Charniak, E. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Workshop on Very Large Corpora*. Pages 161-170. Montreal. Canada.

Harabagiu, S. and Maiorano, S. 2000. Multilingual Coreference Resolution. In *Proceedings of ANLP-NAACL2000*.

Hirschman, L. 1997. *MUC-7 coreference task definition*. Version 3.0

Hirschman, L., Robinson, P., Burger, J. and Vilain, M. 1998 Automating Coreference: The Role of Annotated Training Data. In *Proceedings of the Workshop on Linguistic Coreference*.

Hirst, G. 1981. *Anaphora in Natural Language Understanding*. Springer Verlag.

Leech, G. and Garside, R. 1991. Running a grammar factory: the production of syntactically analysed corpora or "treebanks". 15-32. In Johannsson, S. and Stenstrom, A. (eds.), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton De Gruyter, 1991.

Mitkov, R. 2001. *Anaphora resolution*. Longman (forthcoming).

Mitkov, R., Orasan, C., and Evans, R. 1999. The importance of annotated corpora for NLP: the cases of anaphora resolution and clause splitting. In *Proceedings of "Corpora and NLP: Reflecting on Methodology Workshop"*. TALN'99. Corsica. France.

Orasan, C. 2000. ClinkA A Coreferential Links Annotator. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. ELRA.

Orasan C., Evans R., and Mitkov R. 2000. Enhancing Preference-Based Anaphora Resolution with Genetic Algorithms, In *Proceedings of NLP'2000*, Patras, Greece. 185-195.

Paice, C. and Husk, G. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". In *Computer Speech and Language*. Academic Press.

Passonneau, R. J. and Litman, D. L. 1997. Discourse segmentation by human and automated means. In *Computational Linguistics* 23(1), 3-139.

Poesio, M. and Vieira, R. 1998. A corpus-based investigation of definite description use. In *Computational Linguistics*, 24 (2), 183-216.

Popescu-Belis, A. 1998. How corpora with annotated coreference links improve reference resolution. In *Proceedings of the First International Conference on Language Resources and Evaluation*, 567-572. Granada, Spain.

de Rocha, M. 1997. Supporting anaphor resolution with a corpus-based probabilistic model. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, 54-61. Madrid, Spain.

Tapanainen, P. and Jarvinen, T. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*. ACL.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. 2000. Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000)*. Lancaster, UK (forthcoming).

van Deemter, K. and Kibble, R. 1999. What is coreference and what should coreference annotation be?. In *Proceedings of the ACL99 Workshop on Coreference and its Applications*, 90-96. College Park, Maryland, USA.