Discourse Data in DiET

I. Lewin (3), P. Bouillon (1), S. Lehmann (1/2), D. Milward (3), L. Tanguy (1)
(1) ISSCO - University of Geneva, 54 Rte des Acacias, 1227 Geneva - Switzerland
E-mail : {Pierrette.Bouillon|Ludovic.Tanguy}@issco.unige.ch
(2) DFKI - Saarbrucken - Germany
E-mail : slehmann@dfki.de
(3) SRI - Cambridge - United Kingdom
E-mail : {milward|ian}@cam.sri.com

Abstract

The DiET project provides systematically constructed and annotated test items and associated tools, enabling fast system debugging and evaluation, and automatic linkage from test items to real corpora instances. This paper concentrates on the discourse test suite and its use. The discourse test suite covers discourse phenomena such as pronouns, definites and ellipsis. These can be used to evaluate the coverage and accuracy of implementations of anaphora resolution algorithms. We also examine the text profiling support within the Diet tools. Text Profiling identifies typical and salient corpus characteristics, e.g. the frequency and distribution of part of speech tags and vocabulary richness. Profiling also provides candidate sentences instantiating predefined syntactic phenomena. Profiling enables users to select test-items appropriate to their domain specific corpus. The paper shows how the corpus search engine can be used to identify discourse phenomena in a corpus and presents concrete results of this evaluation scenario.

1 Introduction

The DiET (*Diagnostic and Evaluation Tools for Natural Language Applications*) project[1] has created a comprehensive software package for the construction, annotation, customization and maintenance of structured linguistic data for NLP applications and provides a considerable amount of annotated test data representing phenomena on the levels of morphology, syntax and discourse for English, French and German. The system is designed for use in the evaluation of natural language (NL) applications. It includes support for construction and maintenance of test suites; support for customizing test suites to particular application tasks and support for test suite usage to evaluate particular applications, e.g. helping a user determine that an anaphor resolver is consistently failing on plural definite anaphora. The DiET system is implemented in a configurable, open client/server architecture. A central database system manages the data. A client facilitates data creation and annotation. Multiple servers support customization procedures.

Two sets of tools can be distinguished. The first set allows the user to enter data either manually or by an import function. Data can include sentences, (ordered) groups of test items, phrases, and larger units e.g. paragraphs. The user can describe the data with any annotation chosen from a set of basic annotation modes built on a fixed inventory of data types, e.g. boolean, integer, real, string, tree and marking. Annotation modes are directly associated with display, storage and editing functions. The resulting annotation types can be freely ordered in a hierarchy called an annotation schema. In the system different schemas can be associated with the same set of data, thereby enabling different evaluation studies. Section 2 discusses the DiET discourse test items which have been developed and the discourse annotation schemata.

The second set of tools provides customization facilities which support various tasks: (i) lexical replacement; the user can substitute lexical items in a test suite and (ii) text profiling; this identifies typical and salient corpus characteristics and establishes on that basis a relationship between test suites and domain specific corpora. Section 3 discusses how text profiling can be used and intro-

[1]The project was supported by the European Commission and the Swiss government (Telematics Application Programme LE 4202) and lasted from April 1997 to April 1999.

duces the corpus search engine. Section 4 presents a scenario for evaluating the tool's ability to identify discourse phenomena in a corpus.

2 Discourse Test Items

Discourse test items cover both pronominal and definite anaphora and simple cases of ellipsis. They exist for English, French and German. The annotation schema is the same for all languages, in that feature names are constant across all three languages although the values may differ for different languages. The test-suites themselves are constructed by systematic variation of certain features in the annotation schema.

2.1 Annotation for anaphora

The Anaphora schema has 19 features. Each test-item contains an antecedent and an anaphor. The antecedent and anaphor in each test-item is marked up for: category, sub-category, function, gender and number. In addition, each antecedent is marked up for animateness. Each test-item is also marked up with a phenomenon-name, the locations of the antecedent and the anaphor, a grammaticality indicator (one of: dubious, grammatical, ungrammatical), the number of sentence boundaries between the antecedent and its anaphor, the number of fully valid antecedents for the anaphor (this will be greater than 1 if the anaphor can be linked to more than one antecedent) and the type of the link between the antecedent and the anaphor. The link-type specifies whether or not an anaphor can be replaced by the antecedent without generating unacceptability or change of meaning; and subject to structurally predictable variation such as case-marking (cf. Quirk, 1985, 12.8).

The two remaining features are: the number of expressions which are designated as competing with the actual antecedent of the anaphor and the feature in virtue of which those competitors are not fully valid antecedents. These two features are explained further below.

Example test-suites in French, English and German have been marked up according to this schema. For example, personal pronominal anaphora test-items have been generated by varying gender and number across the syntactic functions: subject, object, indirect object, noun complement and adjectival complement. Examples include:

- *Gilbert* helped. *He* abstained.
- *Cyril* joue. Gilbert *le* cherche.
- *The men* made a proposal. Cyril gave *them* some advice.
- *Ce garçon* joue. Gilbert connaît le nom *du mien*.
- *The proposal* failed. Cyril was averse to *it*.

The test suites do not generally include negative test items e.g. *Cyril joue. Elle rit.* We expect the test-suites to be used for assessing which programs better identify correct antecedents rather than which programs better identify errors. There is a certain amount of implicit negative data available through the use of the competitor features, however, for example *The answer appealed to Marie. It was good.* In this example, *Marie* is a competitor antecedent to *The answer* because it would be fully valid antecedent of *it* were it not for its gender. Gender is therefore the 'competitor feature' in this example.

2.2 Annotation for ellipsis

The Ellipsis schema has 12 features. Each antecedent and anaphor is marked for category, number, tense and 'restrictions'. By 'restriction' is meant a complement or adjunct appearing on an anaphora which is not present in the antecedent (or vice versa). For example, *aussi* and *not* are restrictions on the ellided material in

1 Je voudrais un vol pour Denver. *Aussi* pour Bruxelles.

2 Gilbert's proposal was good. Marie's was *not.*

Each test-item is also marked with a phenomenon-name, locations of antecedent and anaphor, grammaticality and the number of fully valid antecedents just as in the anaphoric test-suites.

In general, ellipsis, and also non-pronominal anaphora, can depend heavily on inference and knowledge. Test-items which depend thus would not be appropriate for our purposes. Consequently, the included data is of simpler textual varieties requiring no (or very little) inference. For example, exclusion of *pour Denver* from the ellided material in (1) is deemed not to require complex inference. Similarly, cases of definite anaphora in the test suites e.g. *One proposal failed. Gilbert was averse to the proposal* are those where the descriptive content of the antecedent and anaphora are identical.

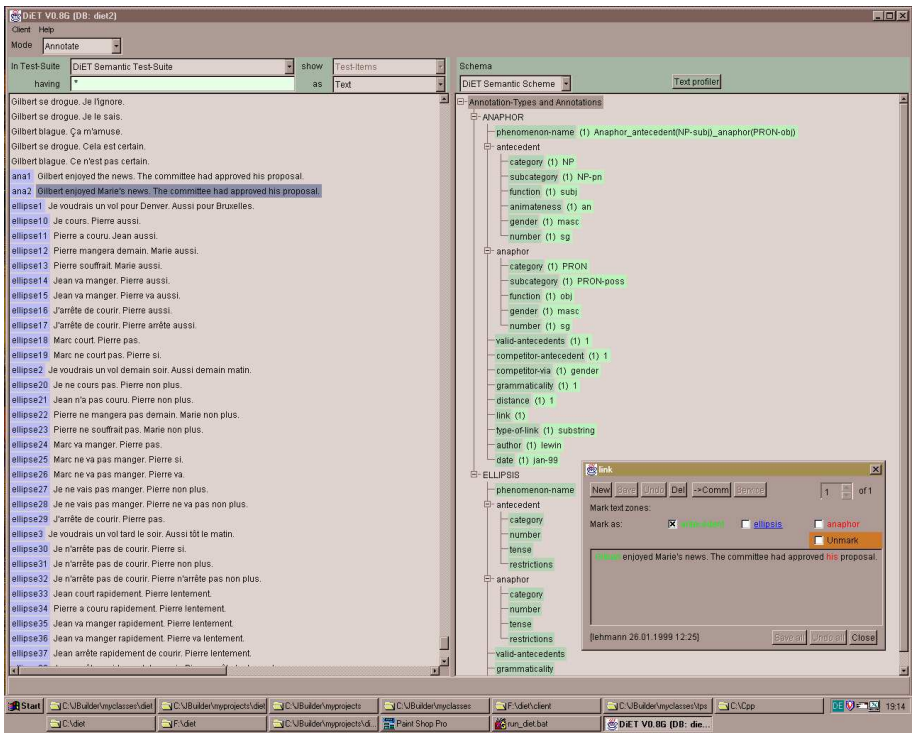

Figure 1: An annotated test item

Figure 1 gives a flavour of how a test item (in this case, an anaphora test item) can be viewed. In the left window a test item is selected. In the right, its annotation is displayed.

3 Text Profiling in Diet

The basic function of text profiling is to identify the typical and salient characteristics of a corpus. These characteristics taken together represent the profile of a particular corpus. This information can be used for various purposes, e.g. providing input for evaluation, weighting a test-item's relevance, helping to map between test-items and corpora, isolating parts of corpora for evaluation.

There are basically two kinds of characteristics. The first one is a set of statistical values extracted from the raw or tagged corpus : frequency and distribution of part-of-speech tags, punctuation, vocabulary richness, etc. These are useful for the user to get a general idea of what generic types of phenomena will be likely to be found in the corpus e.g. question marks and interrogative structures.

The second consists of the number of candidates sentences for a set of predefined syntactic and discourse phenomena. This is obtained through the use of a search engine that relies on the part-of-speech and lemmatization of the corpus. This tool is designed to be used in an environment where a user has large unannotated corpora. We therefore perform linguistic processing only where reasonable accuracy can be assured. In our case we just do tokenisation, sentence splitting and tagging. Noun or verb grouping might also have been possible. Full phrasal analysis (allowing tree search *cf.* TGREP or NEGRA (Skut et al, 97)) would only have been appropriate if users were prepared to manually parse/disambiguate and this was considered unrealistic.

The search engine was specifically developed, following existing ones such as GSearch (Corley et al, 97) or XKwic (Christ 93), and based upon different tagging formats and tagsets. We also developed a graphical user interface in order to help the user design his own patterns, as shown in figure 2.

The pattern language permits patterns over sequences of part-of-speech tags, words and lemmas. Pattern-matching is based on two-level regular expressions: the first level matches a disjunction of words, lemmas and tags; the second level concerns sequences of patterns used to extract sentence structures. More specifically, a sequence can basically be composed from

- An inflected word form;
- A lemma (i.e. non-inflected word form);

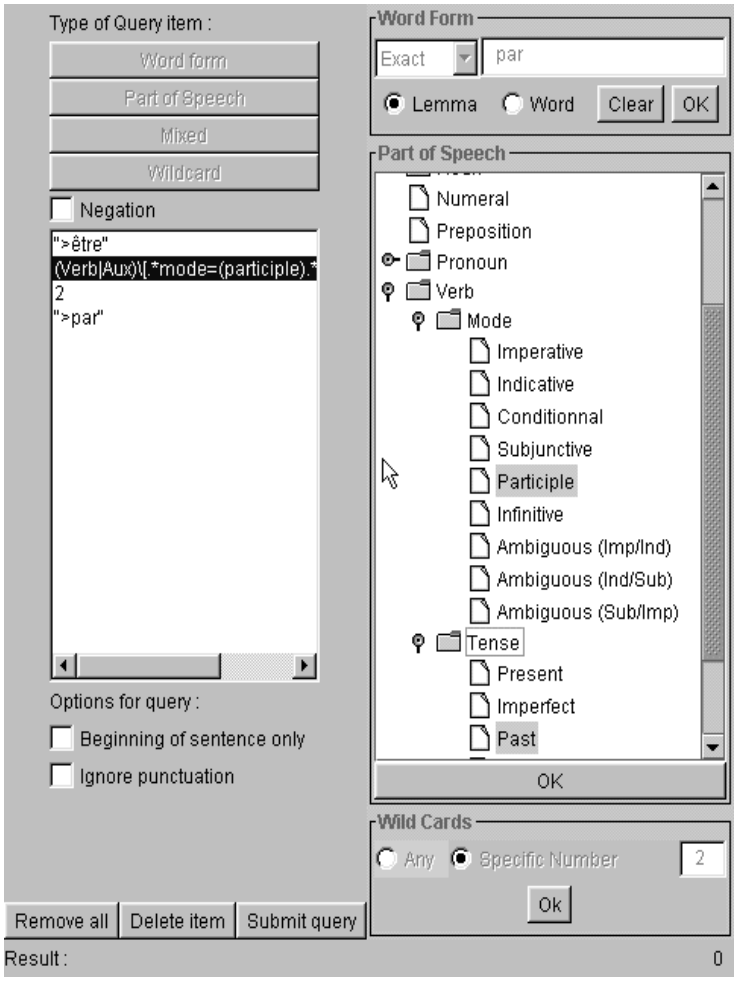

Figure 2: Graphical User Interface for designing patterns

- A part-of-speech tag (depending on the tagset which is used);
- An inflected word form combined with a part-of-speech tag;
- A lemma combined with a part-of-speech tag;
- The negation of one of the above;
- A wildcard (e.g. an integer or '*').

An example of such a pattern sequence is:

*P*ronoun Verb-past Det"one" 3 "car"

This pattern matches any sentence beginning with a Pronoun, immediately followed by a verb (past form), then an occurrence of "one", tagged as a determiner, then a maximum of 3 words, and then an occurrence of "car". It can also be specified that a pattern may only match at the beginning of a sentence. Some patterns can also be designed to be spread across sentence boundaries, when looking for inter-sentence anaphora.

The DiET database also contains predefined pattern sequences for particular phenomena.

4 Evaluation 1: Pattern matching for anaphora and ellipsis

4.1 Objectives

For anaphora, we wish to identify the following phenomena in corpora: the possible kinds of pronouns (personal, possessive, singular, plural, nominative, accusative, etc.); the possible kinds of antecedents (nominal, clausal, etc.); the possible positions of pronouns/antecedents (subject, object,

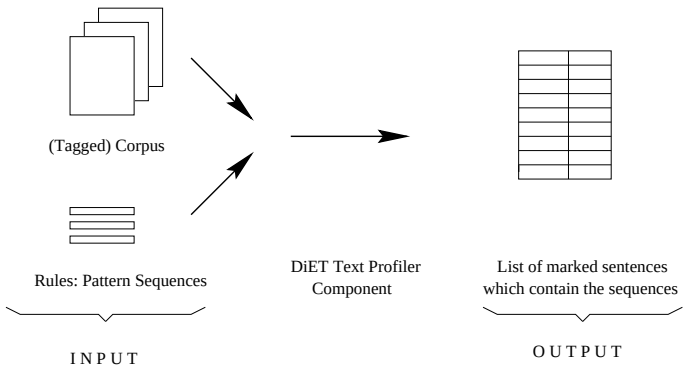

Figure 3: Schema of the text profiler component

complement, etc.).

For ellipsis, we wish to identify: sentences without a constituent (e.g. a verb for example); sentences without a constituent but with a specific restriction.

We present a scenario for evaluating the tool's ability to identify such features.

4.2 A concrete scenario for evaluating the text profiler

Using the text profiler for semantic data requires an annotated text, a set of phenomena that we want to retrieve and a corresponding set of patterns. For this experiment, we use directly the French semantic test suite as the input corpus, since each test item is already identified by a suitably descriptive name which we use to define suitable pattern sequences.

Antecedent(NP-subj)_anaphor(PRON-obj)
Antecedent(P)_anaphor(PRON-compl-ADJ(à))

The test suite allows us furthermore to compare automatically the number of sentences annotated with a specific phenomenon and the number retrieved by the text profiler. To do this, the following is required:

- **Preparation and segmentation** of the semantic test-suite with the MULTEXT segmenter;
- **Morphological lookup:** Each word is annotated by the morphological analyzer `Mmorph` with its base form and its morpho-syntactic description(s) (Bouillon et al., 1998);
- **Conversion of lexical information into syntactic tags:** We distinguish lexical information from syntactic tags. The former is the output of the morphological analyzer: the latter specifies the information that the tagger must disambiguate;
- **Syntactic tagging:** The ISSCO tagger (`Tatoo`, Warwick et al., 1995) chooses the most probable tag for each word from a set of possible tags. The text is thus disambiguated and the original morphological information is restored;
- **Designing patterns:** Each of the 37 different phenomena is associated with patterns;
- **Automatic retrieval of sentences according to patterns and evaluation:** the system computes for each phenomenon the number of actual, correct and expected matches.

4.3 Results and Interpretation

4.3.1 Results

Figure 4 shows an example (for anaphoric indirect object pronouns with preposition à and a sentential antecedent) of the results the text profiler. Precision is the ratio of correct hits to retrieved instances, and recall is the ratio of correct hits to relevant instances. Amongst the expected hits, those retrieved are marked with a star (*). Thus, in this example, out of the 4 existing samples of the `Antecedent(P)_anaphora(PRON-obj-ind(à))` phenomena, the pattern search retrieves all of them but one (recall of 75%).

The results over all patterns/phenomena are shown in figure 5. They lead to two main conclusions: (i) recall is quite good: the text profiler can be used for the identification of specific anaphora/ellipsis in corpora, (ii) precision, however, is poor in the absolute. The reason for that is twofold: First, it is technically impossible to check if an antecedent is correct or not. Secondly, the phenomenon classification is not a partition of the test-suite: most test sentences can indeed

```
Antecedent(P)_anaphora(PRON-obj-ind(à))
Retrieved:          36
Relevant:           4
Correct hits:       3
Recall:             75.00%
Precision:          8.33%
Expected hits :
*Gilbert blague, à quoi je tiens.
*Gilbert part. Je tiens à ça.
*Gilbert part. Je tiens à cela.
Gilbert blague. J'y tiens.
```

Figure 4: Text profiling: Results for a specific phenomenon

Average recall:	79.82%
Average precision:	35.44 %
Phenomena with:	
R = 100%:	23 (out of 37)
R = 100% and P = 100%:	3
R = 100% and P> 50%:	9

Figure 5: Text profiling: General results

exemplify different generic phenomena, as these are hierarchically structured[2]. In the following, problems/limits will be examined in more detail and some some general guidelines for patterns are presented.

4.3.2 Discussion

The errors in the retrieval process can be classified as follows:

Limits of statistical disambiguation. The part-of-speech tagging process is completely automatic and uses a generic statistical language model trained on a large quantity of texts, across different styles. However, some of the phenomena we are looking for in this study are quite rare, and the sentences are generally rather short, which can lead to tagging errors in some cases (approximately 4% in this corpus). Some examples of these tagging errors leading to missed sentences are :

- Gilbert se drogue. Je l'*[Det instead of Pron]* ignore;
- Gilbert viendra. J'y tiens*[Pron instead of V]*;
- Le soleil aveugle*[Adj instead of V]* Cyril. Gilbert le plaint.

[2]In fact, the precision score shown here is a direct comparison between annotation and search hits. It would have been interesting to compute a more complex score, taking all possible phenomena into account for each test item, but could not be easily done due to lack of such information in the annotation schema.

Level of disambiguation As noted above, the final morphological information associated with each lexical item is not fully disambiguated, due to restrictions in the tagset we used. This sometimes lead to a lack of vital information, such as *case* or *gender* for pronouns, resulting in increased noise in the results. One solution to this problem is to enumerate all the possible pronouns in the patterns (see section 4.3.3).

Level of information Our lexicon is restricted to morpho-syntactic information, but some anaphoric phenomena need non-syntactic information to be identified. The most common example is the distinction between *mass* and *count* nouns, having different anaphoric markers in French.

Identifying constituents and their functions As we used no constituent chunking in the text analysis phase, we had to rely on surface analysis to identify sentence parts and their function, e.g. *relative position of the arguments to predicates* (preposed Nouns—Proper Nouns—Pronouns are supposed to be subject, except for interrogative sentences, for example), *occurrences of specific words*, like prepositions for indirect object. Particular syntactic properties like *agreement* can also improve the results as they make the patterns more precise (Lehmann, to be published). Special problems are related to word order, movements in passive and interrogative sentence, and long-term dependencies: all the possibilities have to be taken into consideration in the patterns.

Antecedent identification Most of our pattern sequences cannot be used for identifying the antecedents of anaphora. They only examine the anaphora structure in terms of the type of the pronoun, and the occurrence of at least one possible antecedent (noun, proper noun or pronoun for example). In some cases, the results can be very good. For example, a subordinate sentence contains, by definition, a subordinate conjunction. As a result, the pattern for finding subordinate antecedents are very accurate:

```
Antecedent(subord)_anaphor(PRON-compl-ADJ(de))
Retrieved:          3
Relevant:           3
Correct hits:       3
Recall:             100.00%
Precision:          100.00%
Expected hits :
*Marie pleure quand Gilbert boit. Elle en est
  responsable.
*Marie demande si Gilbert boit. Elle en est
  responsable.
*Marie dit que Gilbert boit. Elle en est
  responsable.
```

However, a clausal antecedent cannot be so easily identified. It may contain other types of antecedents (e.g. nominal phrases, verbal phrases, etc.). Consequently, the precision is very poor:

```
Antecedent(P)_anaphor(PRON-compl-ADJ(de))
Retrieved:          10
Relevant:           1
Correct hits:       1
Recall:             100.00%
Precision:          10.00%
Expected hits :
*Gilbert blague. J'en suis responsable.
```

We finish this section by summarizing some guidelines for designing patterns that emerge from these results.

4.3.3 Guidelines for designing patterns

It is possible to provide a set of rough guidelines which should support the definition of sequence patterns as they are needed as input for the text profiler component. The general rule is to specify the pattern sequences as much as possible with respect to the word form, to agreement and to word order. More specifically, the guidelines can be described as follows (Lehmann, to be published):

- Define the word forms as precisely as possible:

  `Noun-fem-sg` *inst. of* `Noun`
  `Verb-inf ''be''` *inst. of* `Verb-inf`
  `Pronoun ''wh*''` *inst. of* `Pronoun`

- Define the word forms as unambiguously as possible.

  `Verb ''run''` *inst. of* `''run''`
  `Pron ''en''` *inst. of* `''en''`
  (French "en" is ambiguous between a preposition and a pronoun.)

- Add agreement information on the patterns in order to support the identification of selection relations, e.g. agreement information is helpful to identify subject and verb.

- If possible, define patterns which are, if not adjacent, quite close from a word order point of view.

- "Simulate phrases" by indicating an integer or the wild card. If the head of the former phrase is head-final and the head of the latter phrase is head-initial, the two heads can simply be defined as adjacent patters, e.g. an NP followed by a PP can be defined as `N Prep`.

- For phenomena which are difficult to describe by means of a pattern sequence, spell out different flat structures of the possible corresponding sentences.

- Draw a little phenomena hierarchy in order to get a rough idea of the noise (i.e. the additional phenomena) which will be contained in the retrieved output.

5 Evaluation 2: analysing a shallow processing co-referencer

DiET test items not only contain marked up anaphora (which strings are anaphorically related to which other strings) but are also marked up according to features which are relevant to pronoun resolution – for example, gender and number. The DiET tools can be used to extract interesting subsets of this data and used for testing co-reference algorithms. In this way, a more structured picture of the algorithm's performance can be obtained.

This provides an alternative to varying the details of an algorithm itself but running it on the same set of data. In this way, one can also build up a structured picture of the performance of an algorithm over a given set of data. For an example of this sort of an investigation see (Gaizauskas R. and Humphreys, K. 1996). Evidently, such an investigation is possible only if the internal workings of the algorithm itself are available to (and modifiable by) the investigator.

We have used the DiET data to investigate a shallow processing coreferencing algorithm intended to form part of SRI's Highlight information extraction system (Highlight 1999). The shallow parser first detects noun and verb groupings and then uses a) a version of (Kennedy, C. and Boguraev, B. (1996)) to determine pronominal anaphora and b) sortal information and compatibility checks to determine definite anaphora. The investigation using structured data sets extracted from the DiET test suites quickly revealed several areas where the algorithm could be improved or added to. For instance, simply by testing the data with antecedent NPs separately from the non-NP antecedents, it was immediately apparent that the implemented algorithm simply failed consistently on all non-NP anaphora. Also, the tests with pronominal anaphors proved much less successful overall than those with definite anaphors. It was again simple to establish the independence of these two results by simply testing the relative success of definite anaphors against pronominal anaphors on just that subset of the DiET test suite containing only NP antecedents. This revealed precisely the same success ratio for the definite cases and only a small improvement for the pronominal cases. Further tests revealed, for example, that gender appeared not to be a factor in predicting the algorithm's successes and

failures but that subcategory of the antecedent and anaphora clearly was. The coreferencer would never predict an anaphoric link for an antecedent or anaphor appearing in an adjectival complement.

The investigator conducting these experiments was not familiar with either the implementation of the coreferencing algorithm or its intended coverage. Very useful results and pointers to problems were obtained. In order to debug the coreferencer more fully, it was necessary also to examine not only when the coreference was failing but why, by examination of the internal states of the coreferencer. For example, the consistent failure on data containing adjectival complements was found to be an instance of a more general problem concerning any NPs appearing inside a prepositional phrase. In some instances, particular problems were only revealed by inspection of internal states. For example, it was discovered that the tagger was consistently mis-tagging the word "appeals" but the hypothesis that test-items containing that word were generally unsuccessful had simply not occurred to the investigator whilst performing the structured tests.

6 Conclusion

The DiET tools are designed to aid construction, customization and maintenance of linguistic data for use in evaluating NL applications. Discourse anaphoric and elliptical data for English, French and German has been generated and annotated with a reasonably rich set of features. The features are designed to provide useful views of the corpora. We have, for example, used the DiET structured data in testing a pronoun resolution algorithm which uses only a very shallow parsing method. The resulting structured tests were able to guide us quickly to some important features: for example, that the algorithm failed consistently on all non-nominal examples. In other cases, the structured tests allowed us to determine important minimally different pairs, e.g. indicating a lexical entry error (for "woman" compared to "women").

In this paper, we have described the discourse data that has been generated and illustrated the use of another DiET tool, the text profiler for extracting relevant corpus phenomena. The results were encouraging.

For more information about DiET, see `http://diet.dfki.de/` and (Netter 98).

References

- **Bouillon, P., Lehmann, S., Manzi, S. and Petitpierre, D.** *Développement de lexiques à Grande Echelle*, Actes du colloque de Tunis 1997 "La mémoire des mots", AUPELF-UREF et SEVICED, 1998.
- **Christ, O.** *The XKwic User Manual*, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1993.
- **Corley, S., Corley, M. and Crocker, M.** *Corset II User Manual*, University of Edinburgh, 1997.
- **Quirk, R., Greenbaum, S., Leech, G. and Svartvok, J. (1985)** *A Comprehensive Grammar of the English Language* , Longman, 1985.
- **Gaizauskas, R. and Humphreys, K. (1996)** *Quantitative evaluation of coreference algorithms in an information extraction system.* in Corpus-based and Computational Approaches to Discourse Anaphora, editors Botley. S and McEnery T. UCL Press (in press).
- **Highlight (1999)** *Highlight: An information extraction engine.* Cambridge Computer Science Research Centre, SRI International. http://www.cam.sri.com/html/demos.html, 1999.
- **Kennedy, C. and Boguraev, B. (1996)** *Anaphora for everyone: Pronominal anaphora resolution without a parser.* In Proceedings of the 16th International Conference on Computational Linguistics, Vol. 1, pages 113–118, Copenhagen, Denmark, 1996.
- **Lehmann S.** *Towards a Theory of Syntactic Phenomena*, PhD Thesis, University of Saarland, to be published.
- **Milward, D and Tanguy, L.** *Towards the automatic processing of corpora*, 21st meeting of the DGFS, Konstanz, 1999.
- **Netter, K. et al.** *DiET*, LREC proceedings, Granada, 1998.
- **Skut, W, Crenn B., Brants, Th, Uszkoreit, H.** *An Annotation for Free Word Order Languages*, in Proceedings of the Vth conference on Applied Natural Languag Processing (ANLP), Washngton D.C., 1997
- **Warwick, S., Bouillon, P. and Robert, G.** *Tools for Part-of-Speech Tagging.* Technical Report, ISSCO, Geneva, 1995.