DiET in the context of MT evaluation

Judith Klein, Sabine Lehmann, Klaus Netter, and Tillmann Wegst

[1]DFKI GmbH (Language Technology)
Stuhlsatzenhausweg 3, 66123 Saarbrücken
Germany
Tel: (+49) 681 302 5282
Fax: (+49) 681 302 5338
Firstname.Lastname@dfki.de

[2]ISSCO/ETI, University of Geneva
54, route des Acacias, CH-1227 Geneva
Switzerland
Tel: (+41) 22 705 7115
Fax: (+41) 22 300 1086
Sabine.Lehmann@issco.unige.ch

Abstract

The DiET project is developing a comprehensive software package for the construction, annotation, customisation, and maintenance of structured reference data for the evaluation of NLP applications. The DiET system is implemented in a configurable, open client/server architecture with a central database system managing the data and a client integrating construction and annotation facilities and several servers supporting customisation procedures. The application of existing test data to new domains is supported by various means for customisation. These tools allow the user also to establish a relation between isolated, artificially constructed test items and specific corpora by means of a text profiling server, or to harmonise the vocabulary of the test items and the MT system under test by using lexical replacement functions.

Since DiET provides an information system for handling different types of test data for a variety of applications, it can also be employed to support the (linguistic) evaluation of MT systems. Even though the DiET system will contain a substantial amount of annotated linguistic test suite data for English, French and German, the evaluators of an MT system will eventually be in charge of the construction of suitable material for their specific evaluation experiments. The DiET client offers system developers or professional evaluators the possibility to easily construct and annotate their own test data by either choosing the annotation types from the existing annotation type hierarchy or using the configuration mechanism to define new annotation schemata.

1 Motivation

As industrial use of MT technology is flourishing, especially within the expanding international information society, there is an increasing demand for the quality assessment of MT systems in order to:

(i) improve the (linguistic) system performance,

(ii) show if MT technology is suitable for the translation task in a specific application domain (e.g. provide multi-lingual technical documents), or

(iii) compare competing MT systems to inform a potential customer which system is suitable for his/her translation task.

System developers as well as professional evaluators are asked to systematically carry out profound evaluation studies. While final product evaluation clearly has to address various aspects defining the usage of MT applications, such as extensibility, translation speed, or human post-editing efforts, one crucial evaluation dimension for MT systems is the testing and evaluation of linguistic performance.

Effective and efficient linguistic evaluations presuppose sufficient amounts of suitable test and reference material. However, structured and classified test and reference data of various languages – as required for MT evaluations – is often not generally available. This situation is due to the fact that the preparation of such data collections is extremely laborious and time-consuming which makes this task the most expensive factor in carrying out systematic and reliable evaluation experiments. Annotated language resources consisting of (artificially constructed) test suites and (naturally occurring) test corpora serve as input within the testing procedures. The annotations mainly define the (extra-) linguistic criteria illustrated by the chosen test sets.

But the evaluation of diverse applications does not only require clearly defined and described test material: the evaluators of MT systems need, for example, also one translation equivalent or several translation variants which serve as reference for scoring the output of the MT system(s). For them, it is thus even more important to have such reference material at their disposal – although these reference translations are particularly difficult to construct.

In order to record the test runs properly and to provide a complete evaluation report information on the evaluation scenario, such as the conditions under which the evaluation is carried out (black box vs. glass box), the version of the MT system(s) under test, the language(s) tested, the chosen scoring method, the interpretation key, etc. should be registered for each evaluation experiment.

Therefore, the building of test and reference material, the running of tests, the recording of system output, and also the classification and retrieval of evaluation demand an integrated information system where specialised tools support various processes of the evaluation task.

2 DiET Project Aims

The DiET project (**Di**agnostic and **E**valuation **T**ools for Natural Language Applications; http://dylan.ucd.ie/DiET/) aims at developing an information system which supports the task of evaluating NLP applications. The main objective of DiET is to offer a flexible instrument which covers the needs and requirements of very different types of users who wish to employ the system for the diagnosis and evaluation of a broad range of applications. In the design of the architecture of DiET utmost care has been taken to impose as few restrictions as possible on the type of data, type of annotations and external modules for data construction, while still providing a sufficiently useful and extendible platform. The DiET system is implemented in a configurable, open client/server architecture with a central database system managing the data. The usage of a full-fledged DBMS makes it possible to reconfigure and reuse data for different applications in a convenient and flexible way. The central client with a graphical user interface integrates construction and annotation facilities for building annotated test and reference data. Two connected servers support customisation procedures to adapt the data to specific applications or text domains.

Although the main objective of DiET is to provide flexible and user-friendly tools for the construction of test data, the project will also produce a substantial amount of test data in three different languages (English, French, and German). This is quite essential for the project in order to validate the adequacy of the approach and the usability of the tool box on a large and varied set of data.

2.1 The DiET Data Collection

The data collection builds on the TSNLP project (http://tsnlp.dfki.uni-sb.de/tsnlp/) where comprehensive test suites for English, French and German have been developed. For each language about 4500 (grammatical and ungrammatical) test items were built which exemplify a wide range of syntactic phenomena. These test suites will be revised, integrated and modified. Additionally, the database will be extended for morphological test items, for further syntactic phenomena, and for discourse phenomena such as anaphora and ellipsis.

Data construction in DiET follows the test suite approach developed in TSNLP (Lehmann, Oepen *et al*., 1996). Test suites consist of systematically constructed, paradigmatically varied test items, which are controlled with respect to factors such as redundancy and ambiguity. The test items are typically constructed in the form of minimal pairs or sets, with a controlled variation along one or more specific parameters defining a linguistic phenomenon, such that, in the ideal case, a set of items exhausts the logical space defined by these parameters. Such variations often result in *negative* test items, which normally do not occur in corpora but which are nevertheless indispensable for diagnostic evaluation. Depending on the application, *negative* items are not necessarily *ungrammatical*: in controlled languages, for example, a sentence can be grammatical but still "negative" for being outside the scope of the controlled language.

2.2 The DiET Annotation Schema

The DiET system allows the user to create annotation schemata where the values of the individual annotation types, i.e. linguistic, application, corpus, and evaluation-specific attributes, can be drawn from a broad range of data types (string, boolean, number, tree, marking) offered by the system. In the following, an annotation schema is presented which has been worked out within the context of DiET, while section 2.3. explains how to define an annotation schema with the DiET configuration facility.

Annotations assigned to test items serve a range of different purposes, for example, by supporting a more systematic classification of the data. Many of these annotations are also essential for searching and retrieval, i.e. they help the users to extract specific subsets of the data from the database. The annotations proposed in the DiET annotation schema basically describe

- linguistic properties,
- application-specific features,
- corpus-related information, and
- evaluation-specific attributes.

Among the *linguistic annotations*, there are features which interpret and classify a test item as a whole, including language, information about (relative) well-formedness and the test item category (i.e. phrasal, sentence or text). All test items are categorised according to the linguistic phenomenon they illustrate. The annotations specify the name and the description of the phenomenon, its relation to other linguistic phenomena with respect to a hierarchical classification, and the characteristic properties of the phenomenon (e.g. for the syntactic phenomenon *coordination* the features *type* and *number of coordinated elements.*) Other annotations describe the morphological, syntactic and discourse structure of test items. The morphological annotations comprise information on part-of-speech at word level and classify word forms according to uniquely defined morpho-syntactic ambiguity classes. At the syntactic level – following the work in the projects TSNLP (Lehmann, Oepen *et al*., 1996) and NEGRA (Skut *et al*., 1997) – tree representations provide structural information where the non-terminal nodes are assigned phrasal categories and the arcs some grammatical functions (subject, object, modifier, etc.). Annotations at discourse level contain information on the direction (e.g. antecedent) and type (e.g. co-reference) of semantic relations between test item segments.

The *application-specific annotations* can quite trivially assign a test item to a specific application, such that the user is supported in searching for and retrieving suitable test material. However, under this heading one may also find the standard reference or expected output which can be used in the comparison phase of the evaluation procedure. For grammar and controlled language checkers, for example, the error type of ill-formed test items can be encoded and ungrammatical test items can be linked to their grammatical counterparts. For translation systems the test items of the source language can be connected to the corresponding reference translation of the target language.

Corpus-related annotations are those which establish a link between test items and application-specific corpora. This could be information about the frequency (and indirectly also relevance) of the phenomenon represented by an item, or it could even be an index pointing to the occurrences of the respective patterns in the corpus, which can then be used for retrieving such ‘r on those test items is significant which are relevant for the NLP system’s application domain. If, for example, an envisaged corpus contains many ‘real-life’ examples from the text.

Evaluation-specific annotations will help the user to keep track of the set-up of an evaluation and of different test runs, i.e., they will describe the evaluation scenario, including the user type (developer, user, customer), name and type of the system under evaluation, the goal of the evaluation (diagnosis, progress, adequacy), the conditions (black-box, glass-box), the evaluated objects (e.g. parse tree, translation), the examined criteria (e.g. degree of correctness in translation), the applied evaluation measure (the analysis of the system results, for example in terms of recall and precision), and the quality predicates (e.g. the interpretation of the results as *excellent, good, bad*). Since the database is meant to contain evaluation results, the annotations will also be used to record the actual output of the NLP application which is tested. This allows the user to compare the actual and the expected output (stored as reference annotation within the application-specific annotation types).

Besides these item-specific annotations, the database will also contain *meta-information* about the test suite as a whole, such as listings of the vocabulary occurring in the test items, tag sets or generally the descriptive terminology employed in the annotations.

2.3 The DiET Construction and Annotation Tool

The construction and annotation tool constitutes the central application within the DiET system, and operates as a client which can call on different servers. Figure 1 gives an impression on how the annotated test data is represented in the *annotation window* of the system. In principle, the interface aims at simplicity in design and offers clearly arranged operation fields to provide an easy to use annotation tool.

The left window contains the test data themselves which are either test items (e.g. phrases or sentences), (ordered) groups of test items, or segments from test items (e.g. words or even sub-lexical units). Larger units, grouping together related items, can also be defined, including lists of sentences, paragraphs, and discourse or dialogue sequences.

The right window shows the annotation schema with the hierarchically arranged annotation types. If a test item is selected (such as *The teacher might talk to the student, the professor might write him, or the manager might phone him* in Figure 1), the values which have been assigned to the different annotations are shown (e.g. *Test Suite* is the value of the annotation type *Type-of-Testmaterial*). The number in parenthesis after the annotation type indicates how many values the annotation type contains. In the illustration only one value is given per annotation type. Annotations of more complex data types such as the (syntactic) tree appear in a separate window.

The annotation schema is freely configurable: the user can build the annotation type hierarchy in his/her terms by modifying the position of the types, deleting or adding them such that the resulting schema is suitable for the application-specific evaluation study.

The configuration of an annotation schema is very simple. The user is free to choose the kind of test material, as well as the types of annotations associated with the test data. S/he can define new annotation types, which consist of named and parametrised annotation modes, and s/he can organise the types in a hierarchy, which constitutes the annotation schema.

The annotation modes currently implemented cover b*oolean*, n*umber* (integer, real), *string*, *tree* and m*arking.* Annotation types specify them in several respects: they name the mode for its particular use, e.g. *coord-conj, co-reference*, and *syntactic analysis* (cf. Figures 2 – 4), they determine the legal value range, set a list of possible values to choose, and/or provide a default, they say how many values may be entered/chosen and how annotations shall be made at the interface level, they determine whether the user or some service (e.g. a connected part-of-speech tagger) shall provide the annotations,

they select the kinds of objects which the type shall be applicable to (e.g. items or item segments), and they mark annotations as obligatory or optional.

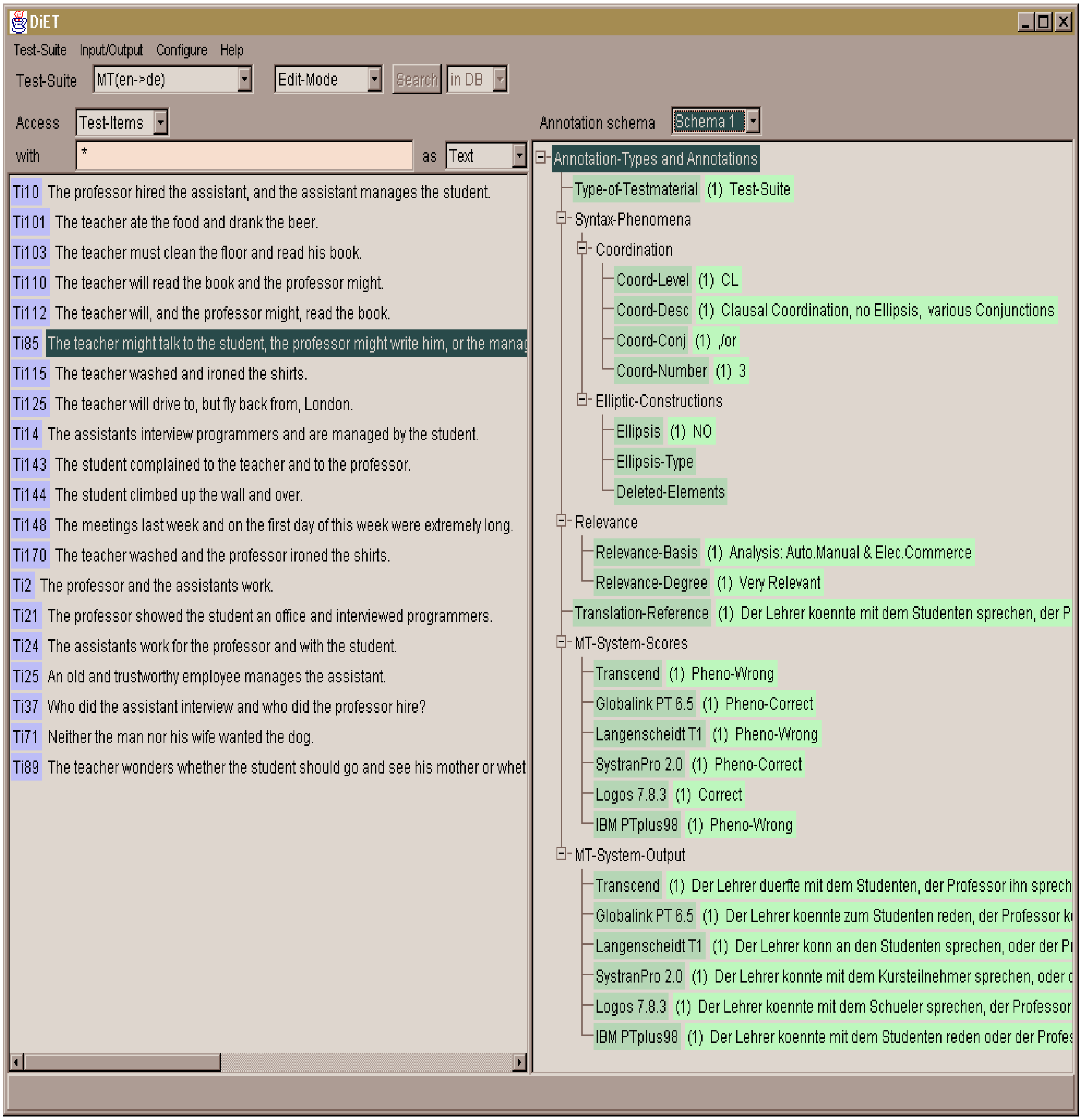

Figure 1 : The DiET Annotation Window

In a highly modular way, the user is offered a choice of basic annotation modes building on a fixed inventory of data types. These annotation modes are associated with the display, storage, and editing functions necessary for handling the data type. This means, for example, that as soon as the user employs an annotation mode *tree*, all the relevant functions necessary to construct and edit tree-like annotations as well as the respective database storage and search functions are activated and available.

Once the annotation types have been defined, the values of the annotations can be assigned to the respective test items. This task is performed in the annotation window environment (cf. Figure 1). First, the user selects a test item by clicking on it, then s/he clicks on an annotation type.

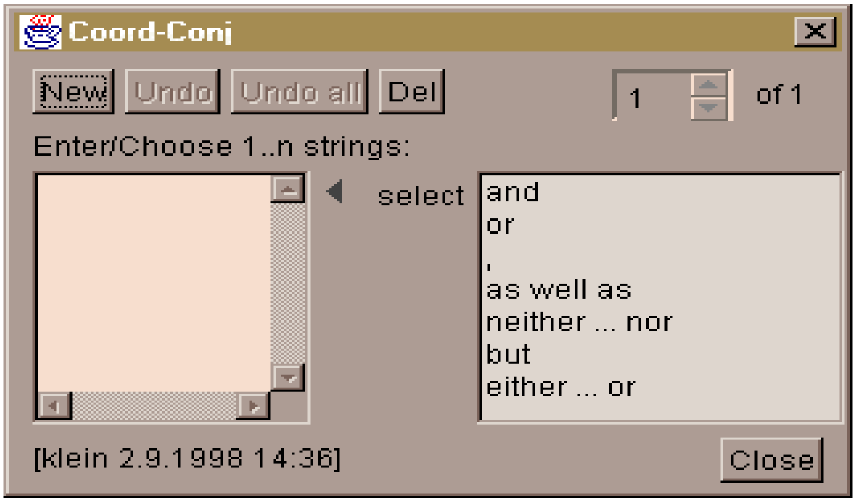

Figure 2: Annotation mode for data type *string*

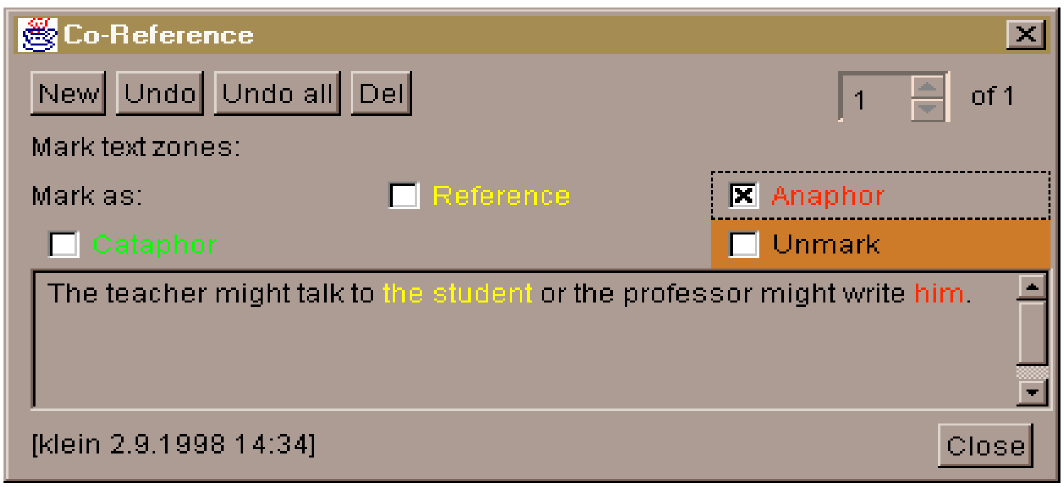

Figure 3: Annotation mode for data type *marking*

The annotation mode interface is displayed whose shape depends on the underlying data type (cf. Figures 2-4). All necessary annotation modes for an annotation scenario can be displayed simultaneously and freely arranged at the screen.

Figure 2 illustrates an annotation mode for the annotation type *coord-conj* (conjunction occurring in a coordinated phrase or sentence). The values of data type *string* can either be selected from a given choice list or alternatively another value can be entered.

Figure 3 shows an annotation mode for the annotation type *co-reference*. The *marking* mode serves among others to annotate discourse relations, allowing the user to define a set of attributes (e.g., referent, anaphor etc.) and to mark-up text segments on that basis.

Figure 4 gives an example of the annotation frame for tree representations. The annotation mode *tree* consists of an elaborate tool which provides a tree drawing facility to build phrasal or relational structures over sentences where both edges and nodes can be labelled.

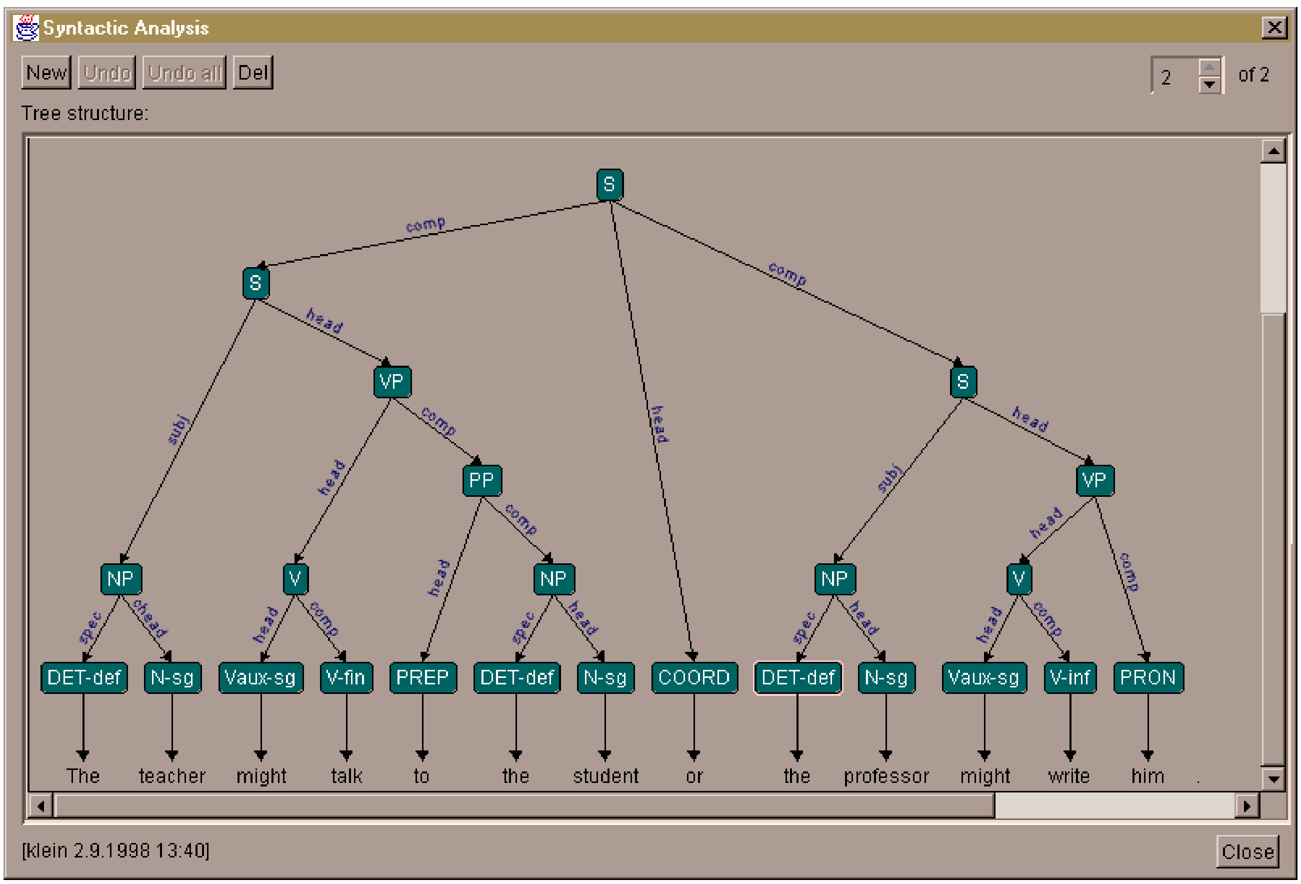

Figure 4: Annotation mode for data type *tree*

2.4 The DiET Customisation Tools

Systematically constructed data are useful for diagnosis but should not be deemed sufficient for adequacy evaluation. The performance of a system should be evaluated on the basis of its intended application. It is therefore very important to know how representative a test item is for a certain (text or application) domain, whether it occurs frequently, whether it is of crucial relevance, etc. For a given evaluation scenario, only performance on those test items is significant which are relevant for the NLP system's application domain. If, for example, an envisaged corpus contains many variants of coordinated elements, the selected test data should also comprise test items illustrating various occurrences of coordination phenomena. Furthermore, the user may wish to cover a specific terminology with the test data in order to better reflect the actual vocabulary within the envisaged text domain.

Thus, it is important to bridge the gap between the isolated, artificially constructed test items and 'real-life', empirically obtained data from corpora, i.e. to relate test suites to test corpora in order to provide weighted test data. Within the DiET system two main tools will offer the means to customise existing test suites to specific text domains and applications: the text profiling server and the lexical replacement server.

2.4.1 The DiET Text Profiling Server

The identification of the typical and salient properties of the texts is what we refer to as *text profiling*. The tools used to identify and classify the corpus characteristics will rely on shallow state-of-the-art corpus processing techniques. These include morphological analysis, part-of-speech tagging, statistical measurements (which can be computed over the entire corpus or only for given localities defined according to a limited set of parameters) and general pattern matching techniques (which are basically used for the extraction of linguistically relevant units). The quality of the result will depend on the success of the shallow processing stage. Accuracy will be much improved if the corpus is already annotated with compatible part-of-speech (PoS) tags, either by hand or by a tagger trained on the specific corpus.

The text profiling tool is based on easily and reliably identifiable corpus characteristics that can be automatically calculated using state-of-the-art corpus analysis techniques. The text profile will contain information about the following types of corpus properties:

- string characteristics,
- lexical properties and
- syntactic information.

String characteristics include format codes, i.e. the identification of types of tags (titles, sub-titles, list items, etc.), their frequency and their distribution. Other relevant string properties contain the orthographic properties of words (e.g. all upper case), the occurrence of alphanumeric sequences and punctuation marks. Information regarding segment types, e.g. paragraphs, sentences and words will also be summarised in the profile.

Lexical properties comprise morpho-syntactic information on word forms and lemmas. Together with information on frequency and distribution of specific lexical units the lexical profile can serve to identify to what degree the lexical data in the corpus matches the vocabulary used in the DiET test items. The information on lexical coverage is needed for lexical replacement, but it could, for example, also be used to classify common vs. less common nouns or give information on type-token frequency.

Syntactic information is based on the part-of-speech tagged words. Lacking syntactically annotated corpus data this tagging will support the identification of phrasal types and syntactic patterns, e.g. finite vs. infinitival phrases or typical PoS sequences indicating specific phrase or sentence types.

An easy to use query language based on regular expressions will allow the user to formulate patterns of various levels of complexity in order to extract specific corpus parts.

The characteristics recorded in the text profile will normally not be sufficient to automatically extract the relevant subset of DiET test items for a given evaluation. The data developed in DiET is designed to provide basic linguistic coverage with a limited vocabulary, often simplified to isolate the different phenomena. Therefore, it would be unrealistic to assume a simple mapping between the annotated test items and the sentences containing these phenomena in the user-specific corpus. A text profile can, however, provide indications as to which items might be relevant and what extensions or modifications of the test data need to be envisaged.

The properties described in the profile may also serve as a basis for assigning relevance measures to the selected test items. If, for example, the text profile records a high percentage of coordinated noun phrases, all test items containing such structures are given a high relevance value. This mechanism relates test suite items to domain-specific corpora and provides weighted test data. It will be up to the human evaluator to assign these relevance measures.

2.4.2 The DiET Lexical Replacement Server

Lexical replacement functions allow the user to harmonise the vocabulary of the test suite items and the user-specific corpus data. Within a given evaluation scenario the adaptation of lexical material may be required for two purposes:

- lexical replacement as a repair: replace words within test items by adequate substitutes, and
- lexical replacement as an insertion strategy: add specific terminology into the test suite.

The repair strategy could be employed if the evaluation experiment is not concerned with lexical coverage. In this case the testing results should not be falsified by poor system results which are caused by unknown words. The simple exclusion of test items containing unknown lexical material is no solution. Given that in a non-redundant test suite a certain phenomenon may only be illustrated once, the simple deletion of a test item from a test set might lead to an inappropriate evaluation. This is so since pertinent phenomena could not be evaluated after such an exclusion. Instead, the lexical replacement routine will replace the unknown word by a word that is both contained in the test suite lexicon and the application lexicon. Equivalence classes for lexical replacement within well-formed and ill-formed items can be built on the basis of a structured lexicon, where categorial and morphological descriptions can be related to each other in terms of type or class identity and generalisation.

In the case of well-formed examples, an unknown word can be replaced by a known word if the unknown word and the substitute share their description. In addition, a replacement is possible if the substitute receives a less general description than the replaced item and therefore only equally or more specific classes can be used for replacement.

In the case of ill-formed examples, an unknown word can be replaced by a known word under the same conditions that applied to well-formed examples. It must be considered, however, that in the case of generalisation, the ill-formedness of a test item could be neutralised.

The insertion strategy allows the deliberate addition of new lexical material into a set of test items. New lexical material might consist of some specialised vocabulary or a specific terminology. DiET will concentrate on providing tools for the first purpose but wishes to extend these tools to cover the second application as well. For a more detailed description of lexical replacement, the reader is referred to Kiss & Steinbrecher (1998).

3 MT Evaluation – an application scenario for DiET

After having described the DiET project in general terms, this part now investigates how the DiET system could be used in the context of MT evaluation. The application scenario is based on the evaluation study carried out by Rita Nübel (Nübel, 1998). The aim of the investigation is to demonstrate that the relevant information required for this concrete evaluation scenario could easily be classified, maintained, customised and accessed using the DiET tool box. Since the DiET system was still under development at that stage, it could not actually be employed to support the evaluation task.

3.1 DiET in the MT evaluation procedure

The evaluation of MT systems is a complex task and requires many different types of information. Since the DiET system offers sophisticated information technology with an underlying database system and various tools to handle different data types, it is well suited to support evaluation. In the following one possible evaluation procedure is sketched. The steps which could be supported by the DiET tool box are listed as bullets, those which must be performed without DiET's help are marked with a star.

- Record the evaluation framework
- Establish a text profile using the DiET text profiling server

* Define construction methodology for MT test data

* Design MT annotation schema

- Define MT annotation schema using the DiET configuration facility
- Insert and annotate test items with the DiET annotation tool
- Extract test sets from the DiET database using the DiET search facility
- If necessary, substitute words of test items by using the DiET lexical replacement server
- Export test sets to MT system using the DiET export facility

* Perform test runs on the MT system using the selected test items

- Import MT system output into DiET system with DiET import facility

* Judge MT system output

* Calculate MT system performance

- Annotate items with MT output information using DiET annotation tool

This papers mainly concentrates on those parts of the evaluation procedure which can be supported by the DiET construction and annotation tool.

3.2 Evaluation Scenario for MT evaluation

First, the user has to define the evaluation framework and specify the conditions under which the MT system should be evaluated. The DiET system allows the user to record this information but at the current stage it is not yet clear how the evaluation framework will be represented.

For illustration reasons, the annotation type hierarchy has been employed to specify evaluation-specific information. Figure 5 presents the concrete setting specified for Rita Nübel's experiment.

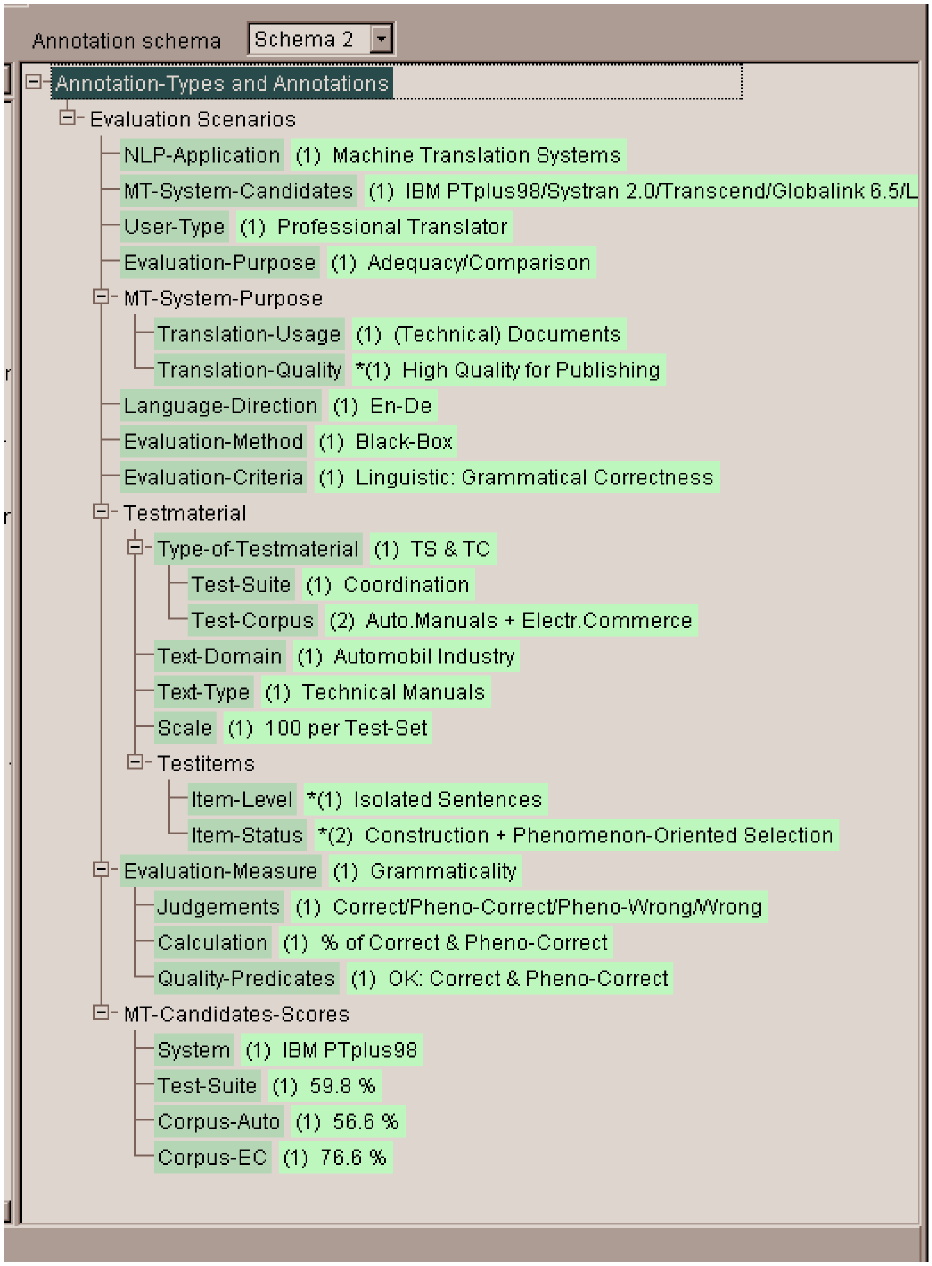

Figure 5: Annotation schema for evaluation framework

The annotation type hierarchy shows in an abstract way, that six MT systems were tested under the perspective of a professional translator who wants to check if the systems are adequate for the given translation and how the six candidates perform in comparison. The task consists in translating English sentences typical for technical document into high quality German.

This linguistic evaluation experiment was carried out under black-box conditions and the examined criterion is the grammatical correctness of the output.

The test material used for the experiment is described in terms of type of test material, text domain, text type, number of test items per test set, and type of test items (e.g. isolated sentences versus paragraphs and artificial construction vs. selection of real corpus sentences). The evaluation measure is concerned with the grammatical correctness of the translations. Each test items is judged as being completely correct, showing the correct or wrong translation of the tested phenomenon, or being completely wrong. The percentage of correct translations together with those translations where the tested phenomenon is correctly translated is calculated. For this percentage the quality predicate *ok* is assigned. Finally the scores of the MT systems on each of the three test sets are given in terms of the percentage of the translations qualified as *ok*.

This description of the evaluation framework is only one illustration on how the evaluation setting could be specified. The user will be free to configure his/her own scenario.

3.3 Annotation Schema for MT Evaluation

Annotating is actually not a task per se, so the data and the choice of the annotations depend very much on what one wants to do with it, e.g. testing linguistic theories, exploiting corpora, or evaluating MT systems. Linguistic evaluations on MT systems require information on the linguistic phenomena tested, their relevance within the envisaged application domain, and ideally, one or several translation equivalents in the target language(s). The actual MT system output – including the scores given to it – should be added as annotation to the source language items in order to easily compare the results to the reference translation and among competing MT systems.

In Rita Nübel's evaluation study on coordination phenomena three test sets of 100 items each were used: one consisting of artificially constructed test suite sentences and two consisting of systematically extracted corpus sentences. All test sentences were classified according to linguistic features describing the coordination phenomena.

The proposed annotation schema exemplified in figure 6 comprises the linguistic, corpus-related, and application-specific annotations specified for this evaluation

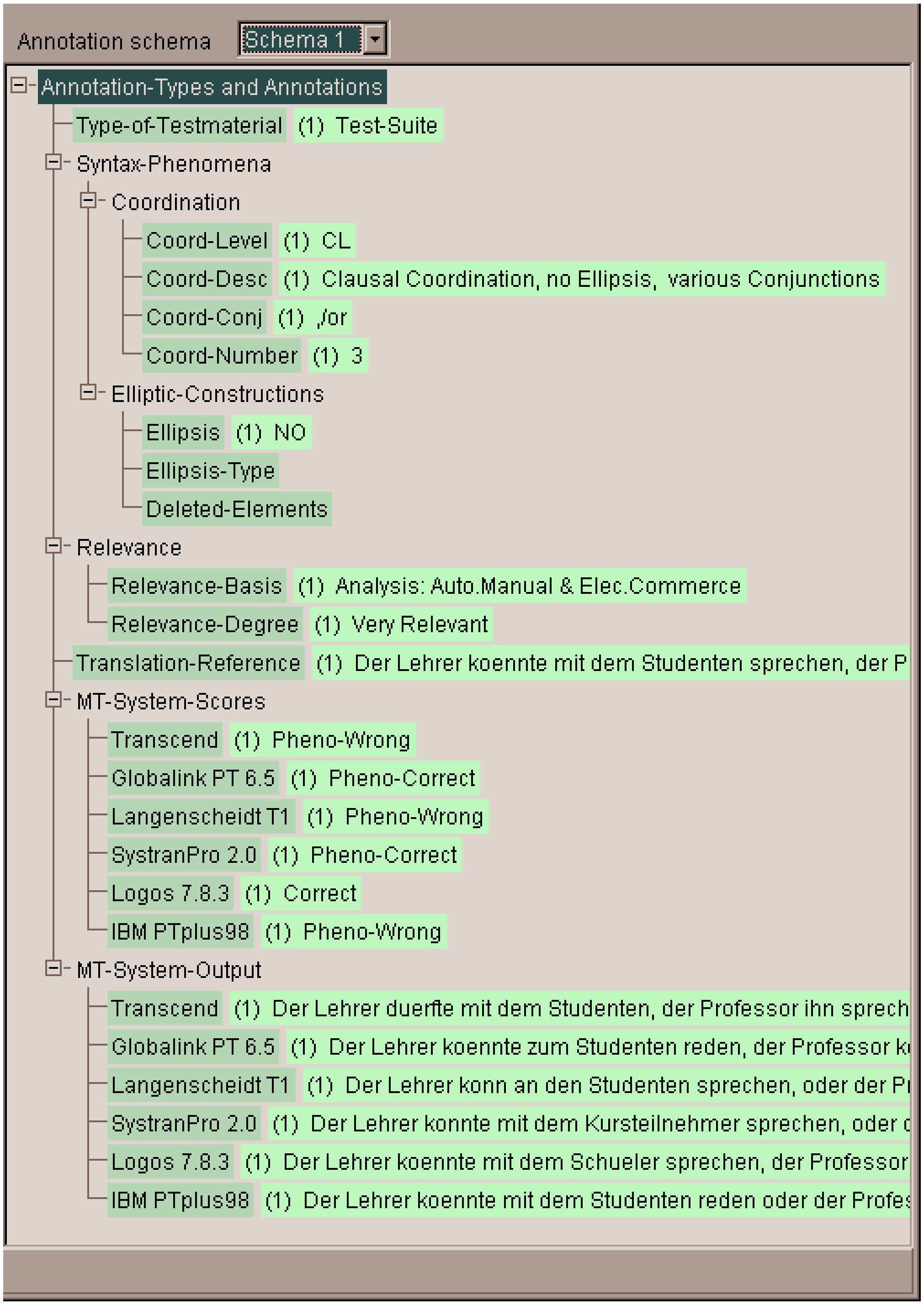

Figure 6 : Annotation Schema for MT evaluation study

experiment. The instantiation of the annotation types show the annotation for a test item illustrating the coordination of three clauses which contain no ellipsis and which are connected by the conjunctions *comma* and *or*:

The teacher might talk to the student, the professor might write him,

or the manager might phone him.

The linguistic annotations try to formalise the above description in an abstract way with the attributes classified under the syntactic phenomena *coordination* and *elliptic-constructions*, namely *coord-level, coord-desc, coord-conj, coord-number, ellipsis, ellipsis-type and deleted-elements*.

According to the frequency of this coordination type in the two test corpora each test item gets one of three possible relevance values (very relevant, relevant, irrelevant). The selected example sentence from the test suite is classified as *very relevant*. This information is included under the heading *relevance* in the annotation schema.

The application-specific information comprises the German reference to the English test sentence, the scoring and the actual German translation of each MT system for the selected test items. The proposed German reference translation is:

> Der Lehrer könnte mit dem Studenten sprechen, der Professor könnte ihm schreiben oder der Manager könnte ihn anrufen.

Under the *attribute MT-System-Scores* are the assigned judgements for the translations: One system delivered a correct translation, two systems provided a correct translation for the coordination part of the sentence and three systems didn't translate the coordinated construction correctly.

4 Conclusions

The main goal of the DiET project is to offer the evaluators (i.e. developers, industrial users and consultants) a reusable tool package to support evaluation. The DiET system gives the user the means to develop test suites consisting of manually constructed structured test items with different types of annotations building on a fixed inventory of data types. However, the tools are not constrained to such types of reference data, but will also encompass the facilities for the treatment and mark up of non-structured textual corpora. Thus the DiET system does not impose any restrictions neither on the test material nor on the annotations associated with the test data.

Since the DiET system can be easily customised to other applications, it is also suited to support MT system evaluation. One possible application scenario for DiET has been illustrated in the context of context of MT evaluation. The test data and annotations are taken from a concrete linguistic evaluation experiment on coordination phenomena. The scenario showed that DiET supports test data classification, annotation and customisation. It demonstrated how an evaluation-specific MT annotation schema could be designed. On the other hand there are some important steps in the evaluation procedure which DiET will not support. The project does not provide neither MT specific, fully annotated test data nor comparison routines to check the actual MT output on any reference data.

Acknowledgement

We thank our colleges from the DiET project Susan Armstrong (ISSCO), Frédéric Joffroy (Aerospatiale), Tibor Kiss (IBM), Bernice McDonagh (LRC), David Milward (SRI), Sylvie Regnier-Prost (Aerospatiale), Reinhard Schäler (LRC), and Ludovic Tanguy (ISSCO) for their contributions. We are also grateful to Rita Nübel from IAI Saarbrücken who has provided us with real-life data from her MT evaluation study on coordination phenomena (see this volume).

References

Kiss, T. et al. (1997) The DiET User Requirements Analysis. D1.1 IBM Heidelberg, 1997.

Kiss, T. and Steinbrecher, D. (1998) Lexical Replacement in Test Suites for the Evaluation of Natural Language Applications. In: Proceedings of the first international Conference on Language Resources and Evaluation, Granada 1998 pp.903-907

Klein, J., Lehmann, S., Netter, K., and Wegst, T. (1998) Construction and Annotation of Test-Items in DiET. In: Proceedings of ESSLLI Workshop on Recent Advances in Corpus Annotation. 10^{th} ESSLLI 1998, Saarbrücken, 1998

Lehmann, S., Oepen S. et al. (1996) TSNLP – Test Suites for Natural Language Processing. Proceedings of Coling, 711—716, 1996.

Netter, K., Armstrong, S., Kiss, T., Klein,J., Lehmann, S., Milward, D., Regnier-Prost, S., Schäler, R., and WegstT. (1998) DiET –iagnostic and Evaluation Tools for Natural Language Processing Applications. In: Proceedings of the first international Conference on Language Resources and Evaluation, Granada 1998 pp.573-579

Nübel, R. (1998) Phänomenspezifische Evaluierung von maschinellen Übersetzungen am Beispiel Koordination. This volume.

Skut W. *et al.* (1997). An Annotation Scheme for Free Word Order Languages. *Proceedings of ANLP*, 88—96, 1997.