Analysis of Existing Test Suites

Dominique Estival, Kirsten Falkedal, Lorna Balkan, Eva Dauphin,
Siety Meijer, Klaus Netter, Stephan Oepen, Sylvie Regnier-Prost.

Report to LRE 62-089 (D-WP1)

Test Suites for Natural Language Processing (TSNLP)

University of Essex

January 1994

Summary

Title	Analysis of Existing Test Suites
Project	LRE-62-089
Document ID	D-WP1
WP	Survey of Existing Test Suites
Date	March 7, 1994
Principal Authors	Dominique Estival
Authors	Dominique Estival, Kirsten Falkedal, Lorna Balkan, Siety Meijer, Sylvie Regnier-Prost, Eva Dauphin Klaus Netter, Stephan Oepen
Reviewers	the consortium

Contents

Analysis of Existing Test Suites

1 Aims of Deliverable

The TSNLP project is aimed at developing a methodology and tools for designing and using test suites for the testing of Natural Language Processing systems. In this project, the first Work Package (WP1) "Survey of Existing Test Suites", has two main goals:

- to identify relevant characteristics for the description of test suites,
- to get an overview of the types of test suites that exist.

Both of these goals share a number of subsidiary aims:

- suggesting some classification among the test data
- finding out the application areas to which test suites have been applied,
- getting some idea of the type of testing for which test suites have been used,
- assessing the reusability of existing test suites.

The two main goals (identification of useful characteristics and description of existing test suites) have been pursued in combination, by refining a list of characteristics serving as a questionnaire for the description of test suites through successive rounds of descriptions of a number of test suites to which the project partners have access.

We are not interested in this Work Package in making detailed statistical comparisons between the various test suites we have analyzed, and we thus do not include the individual descriptions in the report. Rather we want to present a general picture of what already exists in order to build upon it.

This report thus presents and discusses the various characteristics of test suites which we have found important in describing existing test suites, and the characteristics which will be necessary in the methodology for designing test suites.

We first introduce in section 2 the test suites that were available to us and which we have analyzed, then outline in section 3 the methodology we followed in their description. In section 4, we explain the design and purpose of the questionnaire we developed in the course of this Work Package, and in section 5, we give a summary of the results of this questionnaire for the test suites that we had access to. Section 7 then ends this report by mentioning the conclusions that can be drawn from this part of the work, with respect to both the development of the questionnaire and the answers we were able to extract from our study.

2 The Available Test Suites

It must first be noted that, although test suites are often mentioned in the NLP literature, they are in fact neither widely nor easily available. Whether the reluctance of researchers, developers and evaluators to share their test suites is due to company policy or an underestimation of the usefulness of their own test data is not clear. In any case, we were only able to obtain a limited number of test suites (a total of nine at the end of WP1).

A search on the network (on the LINGUIST and LN mailing lists) with requests for contributions and for information resulted mainly in references to the test suites already accessible to the partners, but we still hope to obtain some more test suite data through the contacts that we made this way.

The list on which we settled therefore comprises the four test suites developed by the partners:

- the English Eurotra "ET-UK" test suite (Essex, 1991)
  This test suite was developed at Eurotra Essex in order to test the performance and development of the Eurotra Translation System.[1] It consists of 955 English test sentences, which are designed to test only a restricted number of common syntactic phenomena.

- the English-French and French-English Aerospatiale "AS-NL" test suites (AS, 1991)
  These test suites were constructed in order to test commercial MT systems on performance and adequacy and contain a fair amount of extra-grammatical data. They consist of 1440 English test items, and 346 French test items.

- the French "EGL" test suite (ISSCO, 1991)
  The EGL test suite was based on the well-known HP test suite (see below) and as such intends to provide test data which are useful for the testing of different NLP applications. It consists of 593 French test items and aims at testing the performance of systems and/or their progress during development.

- the German Disco "DITO" test suite (DFKI, 1993)
  This test suite consists of 1540 German test sentences and is being developed in order to provide a catalogue of major syntactic patterns of German, mainly for the purpose of diagnostic testing of NLP systems. The data are annotated with information which goes beyond tagging parts of speech, and are organized into a relational data base. At this stage only a limited number of syntactic phenomena are being covered.

and three other test suites which were available to us at the start of the project:

[1] The notions of *performance*, *development* and *adequacy* will be addressed later (see section 4.1.2) as well as the differences between *checking* and *evaluation*.

- the Hewlett-Packard test suite "HP' (Hewlett-Packard, 1987)
  The HP test suite has been developed to test the performance and progress of NLP systems in general. It contains approximately 1230 English test sentences which cover a very wide variety of syntactic (and some semantic) phenomena.

- the Madeleine Bates test suite "Bates" (BBN, 1988)
  This test suite, entitled "Draft Corpus for Testing NL Database Query Interfaces", contains around 3200 English test items which are queries to a database. In some cases the queries are interrelated, i.e. the second query takes the answer to the first query as input.

- the French-German SYSTRAN test suite, "SYSTRAN" (Stuttgart: Heid & Hildenbrand)
  This test suite has been developed to test verb codings in the SYSTRAN French-German translation system. It contains 853 French test items.

During the course of WP1, two other test suites became available to us:

- the IAI German → French and German → English "IAI" test suites (Rita Nübel, Saarbrücken, 1993)
  This test suite has been developed at IAI Saarbrücken at the request of LOGOS, in order to test their commercial MT system. It contains 2481 German test items and the target languages are English and French. The test suite has been used in a so-called black box evaluation, where the evaluator tries to locate processing problems without being able to access the modules.[2]

- the English "ALVEY" test suite (Edinburgh/Cambridge/Lancaster, 1989/91)
  This test suite is a collection of 1500 English test sentences which were developed to test the Alvey Natural Language Tools (ANLT, a general-purpose morphological and syntactic analyzer).

3 Methodology

The work on describing these test suites was divided into two rounds of descriptions. In the first round, we took as a starting point the draft list of characteristics which had been drawn up during the kick-off meeting for the project (Essex, 6/12/93), and established the first questionnaire.

Each partner used this first questionnaire to describe both the test suite they had developed themselves in-house and another test suite, with the following distribution:

[2] Note that this test suite for LOGOS was not itself available to us, but that IAI Saarbrücken has been so kind as to provide us with the necessary information.

Essex:	ET-UK and HP
ISSCO:	EGL and Bates
DFKI:	DITO and SYSTRAN
AS:	AS-NL and SYSTRAN

Table 1: First Round of TS-descriptions

It was anticipated that this first round of descriptions would lead to the identification of various types of deficiencies in this first list of characteristics: necessary or useful features might be missing, others might turn out to be inadequate, badly formulated, too vague or otherwise difficult to apply in actual practice. A revised and extended list would therefore be elaborated and this new questionnaire would be applied in another round of descriptions. It could then in turn be subjected to further revision and serve as the basis for the study of the design of test suites in WP 2.

Indeed, a number of inadequacies were uncovered, and in the light of comments and suggestions made after the first round, a new questionnaire was drawn up (given in Appendix A) and used in the second round of descriptions of the same test suites. The major amendments in this second version consisted in the addition of a detailed list of syntactic phenomena to check for, introduced to make the comparison of the coverage of the various test suites easier.

For the same reason, the new questionnaire introduced a notation to indicate whether phenomena were tested *systematically* or *at random*, as well as *in depth* or *superficially* (see section 4.1.3 below).

The second round of test suite descriptions then had the following distribution:

Essex:	DITO
ISSCO:	ET-UK
DFKI:	AS-NL
AS:	EGL

Table 2: Second Round of TS-descriptions

In addition, three other descriptions were produced, using the new list of characteristics:

Essex:	IAI
Essex:	ALVEY
ISSCO:	HP

Table 3: Additional TS-descriptions

4 Characteristics of Test Suites

There are several broad ways of classifying test suites:

- monolingual / bilingual / multilingual;
- artificially constructed / naturally occurring data;
- data items at the word/phrase/sentence levels;
- general purpose / specific purpose;[3]
- domain independent / domain specific.

All of these aspects are of primary importance for judging the (re)usability of an existing test suite with a particular NLP system or application, and they are fairly easy to assess. However, in order to provide both useful and adequate descriptions of existing test suites, further characteristics need to be taken into account, some of them reflecting what we want to know about the test suites, others reflecting what the test suites analyzed actually contain. As could be expected, these additional characteristics of test suites are not so straightforward to define, and several of them still need refinement. This work will continue as part of WP2, but the questionnaire that resulted from WP1 already covers a wide variety of test suite characteristics.

In the rest of this section, we discuss the list of characteristics that were found necessary for a useful description of the existing test suites, following the divisions which are used in our questionnaire. The complete list of characteristics making up the questionnaire used in the second round of descriptions can be found in Appendix A. A summary of the resulting descriptions is given in section 5, and a discussion of the insights we have gained in section 7.

4.1 Types of Characteristics of Test Suites

The questionnaire is divided into the following types of characteristics:

[3] The notions of *general purpose* vs. *specific purpose* cover the cases of test suites intended for any kind of NLP systems vs. test suites intended for specific systems or specific application areas (e.g. like grammar checkers, MT systems etc.), specific stages of development (e.g. prototypes vs. commercial systems), or implementations based on a specific linguistic theory, etc.

4.1.1 Specification of Material

This heading concerns available information about the test suite concerning its identification and bibliographic references for it. It includes the name of the test suite, its author(s), the institution where it was developed, the date of writing, and the state of development.

The date of writing and the names of the authors can be very useful information in tracking down the history of the evolution of the test suite and in trying to find potential users to get feed back from them and update our information about the test suite.

In addition, we also ask for information concerning the languages involved, and the number of test items.

4.1.2 Purpose of the test suite

It is not always clear what kind of use for their test suite the test suite designers had in mind, but it is always very important to discriminate carefully between tools meant for the developers of a system and those meant for outside users or evaluators. Under this heading we therefore introduce a number of questions and distinctions which we believe may influence the way test suites are written and applied.

We want to establish what kind of activity a test suite was written for, and whether the test suite is meant to be applied at a particular stage of development of a system. We also need to know whether a specific system type was envisaged, and what kind of users or evaluators were expected to apply the test suite.

First, we introduce a distinction between *checking* and *evaluation*.

Checking refers to the kind of exercise researchers and developers of prototype and pre-industrial systems perform in order to verify the actual coverage of their system at a given moment. The difference between checking and evaluation mainly lies in the interpretation of the results and the actions to be taken based on these but may also influence the choice and construction of test items.

Following the distinction established in the EAGLES Working Group on NLP Evaluation and Assessment, test suites for *evaluation* are further divided into those that are intended to be used for:

- *diagnostic evaluation*, which aims at localizing deficiencies;
- *progress evaluation*, for a comparison between successive stages of development of a system;
- and *adequacy evaluation*, to determine whether and to what extent a particular system meets some pre-specified requirements.

Another important distinction concerns the *type of system* a test suite is designed to test, since this may influence the selection and form of test items. For instance, general purpose test suites may include test items for NPs, APs or PPs in isolation, whereas test items in a test suite intended for question-answering systems will be queries in the form of full questions.

Knowing the type of system envisaged also gives information about the form of the output that the application of the test suite should produce: output ¿from MT systems are typically translations, output from question-answering systems take the form of answers, and output from general purpose NLP systems most often come in the form of tree or graph-shaped representations. The influence of this parameter on test suite design has not been treated any further in WP1 but will be investigated in detail in later WPs.

Yet another distinction concerns the *type of user/tester/evaluator* who is most likely to use the test suite. This parameter may influence the design of a test suite in at least two ways: concerning the *exhaustiveness* of the number of test items given for each phenomenon, and concerning the type of *annotation* chosen for the test data.

In this respect, the important distinction is that between test suites made by and for researchers and developers of prototypes who test their own system and test suites meant for users and outside evaluators with little or no theoretical (computational) linguistic training. In the first case (where testing is often performed on-line and sentence by sentence), a single test item may suffice for the tester to generate further examples for the same phenomenon, and the annotation scheme may use highly specialized linguistic terminology or even in-house jargon. In the second case, a test suite must be as exhaustive as deemed necessary for the testing, and annotations have to be generally understandable, at least in order for the results to be interpretable. Of course such a test suite may also be useful in the first case, and between these two extreme cases lies a whole spectrum of test suite designs and types of testers.

Finally, a test suite may be designed to test a specific stage of development of a system, and that fact may explain a limited or strongly focussed inventory of test items.

The questions about how the test suite was designed, and whether it followed any particular linguistic theory or theory of evaluation are addressed separately.

4.1.3 Coverage

This heading covers a number of related aspects and this section of the questionnaire is the most exhaustive and detailed. Here, we ask about the specification of the *linguistic* and *extra-grammatical* phenomena which the test suite claims to test, with respect to both the *breadth* and the *depth* of coverage of the test suite. It is thus divided into two subsections, the first concerned with the *breadth of coverage* of a test suite, the second with its *depth of coverage*.

To help the respondent answer these questions, this section of the questionnaire consists of an initial inventory of the phenomena which a test suite may cover. This inventory by no means claims to be complete and is only a tool to describe the coverage of test suites in such a way as to render them *comparable*.

Under *breadth of coverage* we list *morphological, syntactic, syntactico-semantic, semantic* and *extra-grammatical* phenomena.

- The list of *syntactic* phenomena included in the questionnaire was compiled from the headings in the various test suites, and follows standard theoretical and descriptive linguistic terminology and practice. Even though the headings are meant to be self-explanatory, in some cases, examples (in French and in English) are given in order to help the test suite users determine the meaning of some linguistic terms with which they may not be familiar.

- The *syntax/semantics* section covers phenomena which may be considered as semantic in nature but which can in some cases be resolved through structural analysis.

- For *extra-grammatical* phenomena, we want to know whether a test suite includes *punctuation*, takes into account the presence of *formatting elements* (such as titles and lists), and deals with all the special *terminography and lexical problems* (typically writing variants) posed by acronyms, abbreviations, proper names, dates, measure phrases, etc...

  We also ask whether the test suite tests for the *robustness* of the system in the presence of *unforeseen input*. Note that this is not the same as the question of terminography, nor nor as the question of ill-formed input (addressed in another section). We would like to know whether the system can deal with unknown words and what kind of side effects such unforeseen phenomena (e.g. an acronym or a foreign word not contained in the system lexicon) have on the processing of the whole sentence where they appear. For example, if the system handles them at all, does it treat all unknown words as nouns by default?

The second part of this section concerns the *depth of coverage* of the test suite, which is more difficult to assess precisely. We ask about the presence of test items for *ill-formed data*, whether the test suite provides for the testing of the *interaction* and *cooccurrence* of phenomena as well as more generally about the *level of detail* and the *exhaustivity* of test data.

We made a number of changes in this section of the questionnaire for the second round, and in answering the new questionnaire we tried to provide this information, but it was not possible to do it systematically, and it still remains difficult to get a good overview of the breadth and depth of coverage evaluated by the test suites.

In particular, after the first round of descriptions, we realized that it was desirable to state for each linguistic phenomenon whether it was tested *systematically* (i.e. intentionally, in isolation, under an explicit heading, with explicit annotations, etc.) or *at random* (i.e. by coincidence: for instance verbal morphology will occur in every main clause and can be a cause of failure when testing another phenomenon), and whether it was being tested *in depth*. This notation allows us to distinguish between test suites which test the same phenomena but not with the same depth.

4.1.4 Presentation

Under this heading, we want to know more about the way the test items are presented, because the presentation of a test suite is an important factor in assessing the *flexibility* of an existing test suite, i.e. in determining whether it is (easily) reusable and modifiable for purposes or applications different from those for which it was designed.

Most important are the two issues of the *division* of the test data according to phenomena, and the *annotation* of the test data.

Division of the data according to phenomena is clearly a minimal requirement for usability, and the more information there is in the section headings of a test suite, the more user-friendly the test suite is. In addition, we want to find out whether and to what extent existing test suites are further divided and whether these divisions are annotated at the phenomenon or at the item level.

As it is part of the TSNLP project to devise an annotation scheme for test data, and tools to manipulate them, we are especially interested in assessing the systematicity and reusability of the *annotations* used in existing test suites. We therefore ask, on the one hand, for a listing of the annotations used (which may suffice to judge whether a systematic scheme is used) and, on the other hand for the existence of documentation which describes design and annotation principles.

As the availability of tools to store and manipulate test data is a very important aspect of the design and use of test suites, we ask about the existence of tools in a separate question.

4.1.5 Characteristics of the Source

The questions under this heading concern information about the source of the test data. We mainly want to be able to distinguish between artificially constructed test suites, in which the range of test items and the vocabulary used is unbiased by any particular text type, corpus or knowledge base, and test suites drawn from specific text types, actual texts or knowledge bases and therefore possibly limited to specific problems and a specific vocabulary.

Test data which have been extracted from real texts may tend to focus on a restricted set of syntactic and lexical phenomena which have been massively observed in e.g. technical documentation and which are known to be problematic for most existing NLP systems.

Answers to these questions will give some indications of possible limitations to the reusability of a test suite in a different context of testing. The perspective of reusability is further pursued by additional questions about the subject area from which the vocabulary is taken, the size of the vocabulary used in the suite, and whether the test suite was designed with a specific application in mind.

4.1.6 Other

In this last section of the questionnaire we have grouped a number of requests for further information, in case it is available. Answers to several of these questions can in fact only be given on the basis of documentation accompanying a test suite or articles describing its creation and use.

We thus ask whether a test suite is designed following some specific *linguistic theory* or *evaluation methodology*, and to what extent this affects the reusability and flexibility of the test data.

Information on *support tools* to store and manipulate test data and annotations, as well as tools for automatic or semi-automatic verification of the results, is, as stated above, crucial for our later work in this project. The questionnaire asks which tools exist, without asking for a more specific description of the tools. Such information should be obtained from the existing documentation.

Finally, we want to know whether a test suite has been applied and, especially, whether available documentation contains an assessment of the adequacy of the test suite for the type of testing/evaluation undertaken.

5 Summary of the Test Suite Descriptions

In commenting the results of applying our questionnaire we follow again the divisions from the questionnaire used in the second round of descriptions (see previous section, and Appendix A). In each subsection, we present the distribution of the test suites for each of the questions, or in the case of the very detailed questions in the *Coverage* section, the generalizations that can be made from them.

It should be clear that the results reported here apply only to the nine test suites that have been examined during the course of WP1.

5.1 Specification of material

For some test suites (e.g. the "Bates" TS), not much information was available and only communicated personally, for others (e.g. "HP") a great deal of information is widely known and published. Nevertheless, the answers given about *source language*, *size* and *state of development* for all nine test suites yield the following picture:[4].

- *source language*: 5 English, 2 German, 3 French.
- *size*: ranges from 346 to 3200 test items. For certain suites, this number is said to be only indicative, and in some cases, different partners have arrived at different numbers when describing the same test suite.
- *state of development*: not surprisingly, no test suite is claimed to be complete. Some test suites must, however, be considered as completed, in the sense of no longer being worked on or used for their original purpose. Thus the questionnaire probably should have distinguished explicitly between test suites still under extension or revision, and test suites no longer being worked on.

Some of the bibliographical references are difficult to find, and it is not clear to what extent they would provide information on either the design or the actual use of the test suite.

5.2 Purpose

Answers to the questions collected under this heading can often not be obtained just by looking at a test suite but depend on accompanying information and documentation. Many of the distinctions we make here for our analysis purposes (e.g. between checking and evaluation, between types of evaluation and between types of users of the test suite) are new and may not have concerned the designers of the test suites we studied, who all had a more or less specific scenario in mind.

The information we have is therefore in several cases insufficient to allow us to state the exact *purpose* of a test suite.[5] This said, the following tentative picture emerges:

- *Checking*: 2 test suites (1 for general purpose NLP systems (DITO), 1 for DB interface (Bates));
- *Evaluation*: 7 test suites (3 for general purpose NLP systems (HP, EGL, ALVEY), 4 for MT systems (SYSTRAN, AS-NL, ET-UK, IAI):

[4]For most purposes, we count the number of the test suites we have studied as 9. The Aerospatiale AS-NL test suite does, however, have both an English and a French source language section, so when we count according to source language we get a total of 10.

[5]We will return briefly to the consequences of this lack of precision in the conclusions.

- *diagnostic/progress*: 6 (HP, EGL, SYSTRAN, ET-UK, ALVEY, IAI);
- *diagnostic/adequacy*: 3 (SYSTRAN, AS-NL, IAI).

For the type of systems, type of testers, and envisaged state of the system being tested:

- *type of systems*: 4 test suites for general purpose NLP systems (HP, DITO, EGL, ALVEY), 1 for a DB interface (Bates), 4 for MT systems (ET-UK, SYSTRAN, IAI, AS-NL);
- *type of testers*: 6 for developers (HP, EGL, ET-UK, ALVEY, DITO, BATES), 3 for system users or external evaluators (SYSTRAN, IAI, AS-NL);
- *state of the system*: 6 for systems under development (HP, DITO, Bates, EGL, ET-UK, ALVEY), 3 for commercial systems (SYSTRAN, IAI, AS-NL).

Considering *material* and *purpose* of the test suites together, we can summarize the results as follows:

- half of the test suites (5) take English as source language;
- at least 7 out of 9 test suites have been used for evaluation (as opposed to checking only);
- a diagnostic purpose is involved in all cases (i.e. for both checking and evaluation, for the latter with or without progress evaluation or adequacy evaluation as secondary purpose);
- 4 out of 10 test suites aim at testing MT systems, 4 are for general NLP analyzers, 1 is for a database query system
- in 6 out of 9 cases the test suites are designed for system developers
- 6 out of 9 test suites are designed for systems under development

Most of these results fit in well with our general expectations concerning design and application purposes for existing test suites. The notable exception here is the fact that half of the test suites we have collected address MT systems. We are, however, aware of the existence of several other test suites, and it would be premature and probably false to conclude that test suites are especially used to test MT systems. Rather, it could be a perverse effect of the composition of our group and of our professional history.

5.3 Coverage

A. Breadth of coverage

5.3.1 Morphology

No test suite in our set actually tests the full *inflectional* morphology of a language systematically. Most typically, it tests whether all tenses are recognized in a verbal paradigm, but not whether all regular paradigms are covered, nor whether irregular forms are correctly analyzed. *Derivational* morphology is not treated by any of the test suites studied, and only 1 (ALVEY) includes some tests of *compounding.*

5.3.2 Syntax

As we said above, this is the most extensive and detailed part of the questionnaire. Although giving a framework for this part of the description seemed easy at first, taking existing headings as guidelines and using common linguistic knowledge to augment them as needed, this turned out to be surprisingly challenging and this is a part of the task which was found to be quite difficult by most partners when doing the various descriptions.

This is mainly because few test suites actually use the same terminology or even divisions for the linguistic phenomena being tested. One exception is the "EGL" test suite for French, which was designed after the "HP" test suite for English and thus mostly follows the same patterns.

The test suites that are meant to test a complex system (such as the "Bates" test suite, which tests a particular question answering system) in fact difficult to compare with test suites that are meant to evaluate the grammar component of any NLP system under development (such as the "HP" test suite). For example, almost half of the "Bates" test suite thus consists of "Queries organized not by syntax but by activities often performed in a database retrieval task". A test sentence like "Tell me how many ships we have" is here listed under the heading "Counting, totals, size of sets". Re-classification of all of these queries in terms of grammatical phenomena (which would be necessary for an in-depth comparison of the syntactic coverage of test suites), would be a task going well beyond our purposes in this WP.

Obviously, there is also a clear difference between a test suite which has been designed to test the range of syntactic phenomena covered by a specific system under development (e.g. "ET-UK") and one that is meant for general purpose (e.g. "HP"), but the main difference for us here will be that the first one can omit phenomena which the second will have to cover.

To summarize the results:

- all test suites provide for some questions and treatment of negation:

- only 5 test suites provide for some imperative constructions:
- only 3 test suites provide for some conditional constructions:
- 5 test suites treat sentential modification with adverbs; only 2 treat sentential modification with participial clauses;
- all test suites treat the argument structure of verbs and verbal subcategorization;
- most test suites (8 out of 9) focus on the structure of NPs and treat complex determiners; 6 treat the modification of NP with adjectives and relative clauses; only 3 treat NP complementation;
- while all test suites address infinitive and clausal complementation, only 6 test suites treat predicative constructions, 7 treat prepositional complements (addressing the issue of the choice of preposition) and only 6 treat pronominal constructions and pronominal complements;
- all test suites have extensive treatment of the verbal group and address the issues of auxiliary verbs and the placement of negation and adverbs.
- 7 test suites treat AP complements, usually comparative and superlative complements, and 5 treat the modification of AP with adverbs;
- 8 test suites tackle the problems of coordination, mostly at the sentential level, but 6 also address the issues raised by agreement in coordinated NPs or VPs; only 2 look at extraction from coordinated structures;
- only 3 test suites have some test items for parentheticals, and 1 for reported direct speech;
- few test suites (at most 2 for each of these constructions) treat Dislocation, Topicalisation, Cleft, Pseudo-cleft or Extraposition.

The sub-section *Idiomatic expressions* is meant to cover the recognition of syntactic units (not compounds) larger than the word. Only two of the test suites in our pool actually contain such test items.

5.3.3 Syntax/Semantics

The semantic phenomena which can be resolved through structural analysis typically constitute areas which a test suite will either ignore or try to treat in depth. However, no test suite actually included many of these. We found some cases of:

- scope ambiguities: 3 test suites with scope of negation, 3 with scope of adverb, 2 with quantifier scope, and 3 with PP attachment.

- anaphora: 4 test suites.
- adverb co-occurrence restrictions: 2 test suites.
- negative concordancy (and Negative Polarity Items): 5 test suites.

5.3.4 Semantics

Including this section of questions was wishful thinking, as most test suites actually do not work at this level, beyond the semantic phenomena which are treated in the *Syntax/Semantics* section.

One notable exception is the "Bates" test suite, which specifically tests for ambiguity and for correct understanding of queries.

The "HP" test suite contains some inter-sentential anaphora, of the kind that are most easily resolvable by appealing to heuristics (e.g. "A pronoun is most likely to be coreferential with an NP in the same syntactic position in a preceding clause"), but do not require any semantic processing.

Some semantic processing is often used by NLP systems in disambiguation of e.g. verbal polysemy but we found no instances of tests for this type of semantic phenomenon in the test suites available to us. It is however claimed that the "IAI" test suite contains some tests for selectional restrictions placed by verbs on their arguments.

Even for test suites which explicitly include test items for semantic phenomena, we do not know how the interpretation and acceptability of these test items is to be assessed. In the case of the "Bates" test suite, one clear measure of success is whether the query is answered successfully, but for other test suites, we cannot know what it means for a system to succeed on a sentence exhibiting scope ambiguity for example. This will be an area of research for WP2.

5.3.5 Extra-grammatical phenomena

Standard punctuation is present in most test suites, as well as proper names presumed to be part of the system lexica. Apart from this, only the "AS-NL" test suite takes into account formatting phenomena like itemized lists and formulae.

"AS-NL" (and to a certain extent also "Bates") also includes phenomena like complex date expressions, currency phrases, abbreviations, company and organization names, and acronyms. All of these are given in different writing variants (e.g. both "AECMA" and "A.E.C.M.A.") which will not all be coded in the system lexicon, thus providing testing also of the robustness of a system in the presence of unforeseen input. For this latter purpose, the "AS-NL" test suite also includes test sentences containing foreign words.

B. Depth of Coverage

5.3.6 Ill-formed Data

Under this heading, we want to know whether ill-formed data is tested, and whether it is tested systematically. Most test suites (7 out of 9) test ill-formed data, but although some of them exhibit a certain degree of systematicity in testing some types of ill-formedness, none can claim to be absolutely systematic in this. One of the conclusions of this study is actually that no test suite could ever be systematic in this respect, because there are too many ways to derive ill-formed test items from a well-formed one.

Test items for ill-formed input are generally not annotated, beyond the "*" widely used in linguistics, so their presence presupposes a knowledge of the linguistic constraints that have been violated. This might be quite difficult for testers who are not trained linguist, even in their native language.

5.3.7 Interaction and Co-occurrence

It turned out to be difficult to evaluate the level of interaction in the test data. The best that could be done was to note when interaction is purposefully avoided, and when it is explicitly intended. Most of the time, interaction between phenomena is implicitly avoided, (the "Bates" test suite once again being the exception), but of course, all full sentences will test for several interacting phenomena such as agreement and verbal morphology, at the very least.

5.3.8 Level of Detail and Exhaustivity

This was also a very difficult question, which it has not been possible to answer adequately. The questionnaire covers most of the syntactic core phenomena and can be used to assess the breadth of coverage, but better formulations for criteria to describe levels of detail must be worked out in the next phase of the project to judge if a phenomenon has been covered in any depth.

5.4 Presentation: Logical and Physical

As the phrasing of the questions in the questionnaire left notions like *annotation, logical, physical* and *user-friendly* undefined, these were open to different interpretations and the answers given to these questions are not consistent. These notions will have to be refined for future work, especially in WP2.

Nevertheless, the following picture emerges:

- *Annotations*: All test suites have some kind of annotation but only 1 ("DITO") uses a systematic and fine grained annotation scheme. For each test sentence is thus indicated:

  - well-formedness
  - length
  - area of phenomenon
  - position of finite matrix verb
  - position of nominal and prepositional phrases
  - positions of adjective phrases and adverbials
  - position of sentential and infinitival complements
  - position of relative sentence

  Further annotations are provided for verbal government data, coordination data, semi-idiomatic expressions and relative clauses.

- The remaining test suites group sets of test sentences under headings and subheadings. As already noted above, only 2 test suites ("HP" and "EGL") use roughly the same section headings, the others use classification schemes which mainly differ in degree of subdivision (i.e. one test suite has a heading *Infinitives*, others distinguish between *Subject Control Complements, Object Control Complements* etc.). As none of these test suites exhibit a systematic annotation scheme of the kind we envisage, we have abstained from any in-depth study of their respective merits and shortcomings.

- *Documentation*: 8 out of the 9 test suites have known documentation, in the form of at least one available article. A useful question would have been whether and to what extent the information given in these articles would enhance the reusability of the test suite, but we did not formulate any further questions about the issues treated in the documentation, assuming that this would become clear from answers given to various questions throughout the questionnaire. Answers to this question are therefore just given in the form of pointers to the bibliographic references. We will return to the question of documentation in our conclusions below but should note here that "DITO" is exceptional in having specific documentation on data annotation because it is designed to allow easy manipulation of the data.

5.5 Characteristics of the source

As to *source of the test data*, including *type of vocabulary* and *size of vocabulary*, the distribution of the test suites is as follows:

- *source of the test data*:

  for 6 test suites, the data is artificially constructed;

for 2 test suites ("AS-NL" and "ET-UK") the data is partly derived from sentences occurring in a specific corpus;

for 1 test suite ("Bates"), the artificially constructed test items are meant to represent typical queries for the particular application database.

- *type of vocabulary*:

  4 test suites are built with general vocabulary;

  2 test suites purposedly limited the vocabulary to office or personnel management ("HP" and "DITO");

  for the other 3 ("AS-NL", "ET-UK" and "Bates") the vocabulary is largely determined by the corpus or application (see above).

- *size of vocabulary* (estimation):

  for 2 test suites ("Bates" and "IAI") it is too large to estimate;

  for 3 test suites ("AS-NL", "SYSTRAN" and "ALVEY") it is estimated as 800 items or not much larger;

  for "DITO", as 650;

  for "ET-UK", as 400;

  for "HP" and "EGL", as no more than 200.

This information alone is of course insufficient to judge the potential reusability of a test suite but it can be observed that the two artificially constructed test suites which are meant to be general purpose ("HP" and "EGL") constrain themselves within a highly limited vocabulary. This should at least minimize the efforts needed to render any system lexicon compatible with the vocabulary of the test suite. At the other extreme, system and application specific test suites like "Bates"and "IAI" seem less immediately reusable except for highly similar systems.

5.6 Other

We wanted to know whether the test suite had been designed with a particular linguistic theory in mind. None of the other test suites appears to be biased towards a particular *linguistic theory*, and certainly not to the extent that it could not be used for testing a system designed within another theory. Even though the "ET-UK" test suite follows to a certain extent the EUROTRA framework for which it was designed, its linguistic headings are general enough that they are not biased towards any linguistic theory and could be reused for other systems.

Actually, none of the test suites studied were found to depend on a specific linguistic theory, but we can observe that some annotations used as section headings (e.g. "Paycheck sentences" and "Bach-Peters sentences" in the "HP" test suite) require a certain level of sophistication in theoretical linguistics since no further explanation of why such sentences are problematic is offered.

As to adherence to some specific *evaluation theory* or methodology, only one test suite ("SYSTRAN") claims to have followed the guidelines given in King and Falkedal (1990). In addition, the "HP" test suite of course itself was trying to establish the foundations for design and use of test suites, on which "EGL" was later based.

The remaining test suites have all, as far as we can tell from the documentation available, been created following each their own testing or evaluation strategy.

Only one test suite ("DITO") was found to be associated with a set of *support tools* comprising a relational database manager, a graphical database browser (under development), extraction tools, consistency checkers and vocabulary administration tools. The "HP" test suite is also reported to be supported by certain tools but no documentation on these are available to us at the moment.

The question of whether a test suite is flexible depends heavily on the annotations of that test suite and the available tools. As in many cases only a very vague notion of annotation (i.e. section headings) was present, and hardly any tools were available, most test suites can not be defined as flexible. Positive exceptions to this seems to be the DITO test suite, and maybe to some extent HP. It remains, however, an open question what exactly the requirements are for a flexible test suite.

The usefulness of test suites is clearly enhanced by tools going beyond the availability of a good editor and providing at least easy means to perform lexical substitution, but the discussion of which are the most useful did not go very far, because very few test suites actually include such tools. "DITO" does, since it is part of an environment for using test suites, and "HP" is reported to have some, but there is no documentation on them.

Information about the actual use of the test suite and results from users is in many cases not available or very scanty. Direct results are presented in the MT-oriented "AS-NL" and "SYSTRAN" test suites, which have versions that list resulting translations together with the input sentences. For the "AS-NL" test suite, as it was applied to two MT systems, some of the results are given for both systems.

We were especially interested in whether available literature on a test suite contained more principled discussion of its usefulness of the test suite in the light of the results obtained. Such discussion is reported for the "HP", "SYSTRAN", "AS-NL" and "ET-UK" test suites, but this is one aspect of this study which proved extremely disappointing. One conclusion of our study is that it is virtually impossible to assess the usefulness of a test suite without having access to results or at least discussion of the results of using the test suite.

6 Lessons from using the questionnaire

In the course of our description work, the interrelation between some characteristics became clear very early. For instance, one aspect that makes existing test suites reusable and useful for testers other than the test suite developers is the *existence and availability of documentation.* At the same time, external documentation and annotation schemes in the test suite constitute also key factors in determining the *flexibility* of a test suite. On the other hand, flexibility is also affected by the *level of linguistic knowledge* that the user is assumed to have.

Obviously, certain characteristics have more importance than others, but these are not always the ones most easily described. *Flexibility* for instance is clearly an extremely desirable feature for a test suite, but it has proved very difficult to characterize this notion, and a more refined definition remains to be elaborated in WP2.

One characteristic which we expected to have an impact on the design of a test suite was the test suite's *purpose*, particularly concerning the differences between test suites designed for diagnostic and those designed for adequacy evaluation. However, of the three test suites used for adequacy evaluation for which we have descriptions, only one presents notable differences from the others. The main difference resides in the testing for extra-grammatical phenomena and in the fact that this test suite concentrates on the phenomena which are expected to cause problems. In addition, this test suite does not aim at covering a full range of syntactic phenomena and it ignores those which are unlikely to occur in natural text or which were considered unproblematic.

This finding suggests that some of the issues relevant to *adequacy evaluation* are often ignored by the designers of test suites. Nevertheless, although test suites are primarily considered as a tool for system developers and thus tend to concentrate on the syntactic coverage of a system, the adequacy issues are of importance as system developers normally intend to bring their products to the market and then syntactic testing alone is not enough. For that matter, the usefulness of all test suites would be greatly improved by the inclusion of extra-grammatical test data allowing the developer to test the system's behaviour with regard to acronyms, abbreviations, punctuation, formatting, etc.

Another difficulty was encountered when trying to judge the *breadth and depth of coverage* of a test suite. Both breadth and depth of coverage of a test suite are of great importance in describing its linguistic aspects. Of course, it is often the case that one is chosen at the expense of the other: either few phenomena are tested in depth, or a number phenomena are treated very superficially. In fact, most test suites test a few phenomena in depth and have only a few test items for a number other aspects.

One omission in our questionnaire concerns the question whether test items consist of sentences, parts of sentences or sequences of sentences. In fact, only rarely do the existing test suites go above or below the sentence level, except for a few items for inter-sentential anaphora in the “HP” test suite and in sequences of queries in the “Bates” test suite, and in some phrasal (mostly NPs) items in most of them. In addition, reflecting as it does the test suites that were examined, the questionnaire is not very strong on semantic or pragmatic phenomena. This is in fact probably also an accurate reflection of the state of the art in general linguistics.

Other characteristics and properties of test suites which need to be taken into account in the design of test suites in WP2 will include the control of interaction between phenomena, control for frequency of the phenomena in corpora, and the multilingual, or at least bilingual, dimension.

Nevertheless, after having used and refined our questionnaire, it has become possible to evaluate the range, shortcomings and strong points of existing test suites. Some of these points will already have emerged from the above discussion, but we can summarize them as follows:

- Test suites generally concentrate on checking syntactic coverage, leaving other linguistic areas such as morphology, semantics, pragmatics and discourse aside.
- Few test suites include extra-linguistic phenomena such as acronyms, abbreviations, punctuation, formatting.
- Test suites are generally constructed to test NLP systems under development, for evaluation by their developers.
- In most cases the design of test suites has been ad hoc, except for those, like the “HP” test suite and those modelled after it, or the “DITO” database, which try to aim for theoretical neutrality and exhaustivity.

The importance of proper annotations in a test suite has already been mentioned several times. Indeed, it is not quite clear of what value a test suite would be which did not indicate what is being tested. Such a test suite could only acquire some value from the evaluator using it: for a skilled computational linguist the exercise will be a lot more informative than for an untrained user. For an outsider, the test suite would produce much less informative results, since without annotations it is not always clear why a certain test item is included in the test suite and what conclusions can be drawn from the test output.

This may actually be one of the reasons why most of the test suites in our collection are so little used. Most of them have only been used by the groups involved in their development but not by outsiders. Other developers often seem to decide that it is more useful to compile their own collection of test data, not just because they can tailor them to their system and to their notion of what the system should cover, but also because the results will be easier for them to interpret. We do not consider this situation desirable and one of the goals of this project is to try to prevent the multiplication of effort that we have witnessed in test suite building. A good annotation scheme which allows a better understanding of the test data and the interpretation of results is therefore of the utmost importance.

7 Conclusion

The main purpose of the test suite description exercise undertaken in WP1 was, as stated in the beginning of this report, to identify relevant characteristics of test suites and to get an overview of the type of test suites that exist.

As it is the goal of the whole TSNLP project to provide and test a methodology for designing test suites, with an emphasis on annotation schemes and on tools to simplify the construction and reuse of test material, we focussed our study on three issues:

- the purpose of a test suite and the parameters or factors to take into consideration;
- systematic annotation schemes, support tools, documentation and questions of reusability and flexibility of test data;
- the type and amount of test data, and questions of depth, breadth and exhaustivity of coverage.

In addition to the more detailed results described in the previous section, we can here summarize our main findings as follows:

- *Documentation*: Quite generally, existing documentation on the test suites we have studied is insufficiently informative on issues of purpose and experience of application. This may be explained by the fact that most of the test suites have only been applied by their developers and only to the system they were developed for. However, the usefulness of studying existing test suites is severely limited when their purposes and whether they have proven adequate both remain unknown.
- *Purpose*: The influence of the various aspects of purpose on selection of adequate test data and on requirements on annotation schemes in the interest of flexibility and reusability of test data is largely an unexplored area which it is hardly possible to explore on the basis of our set of test suites.
- *Tools and annotation schemes*: Only one test suite ("DITO") is associated with a systematic annotation scheme, support tools and documentation for them. The test suite and the tools are still under construction and it has not yet been applied, but as the developers of "DITO" are partners in TSNLP we will naturally be able to keep informed on what seems to be the most promising existing approach.

- *Coverage*: Some of the test suites do contain sections with systematic and elaborate sequences of test items for selected phenomena but we found no attempt at characterizing the depth or breadth reached or aimed at.

Given all this, a final conclusion is that the tasks delineated in the TSNLP project are indispensable in order to make test suite testing a viable approach and not just an attractive idea.

8 Bibliography

References

[1] D. Arnold, R.L. Humphreys, L. Sadler, "Evaluation: An Assessment", in *Machine Translation* Special Issue on Evaluation, Vol 7, in press.

[2] Bates, Madeleine. (1988). "Draft Corpus for testing Natural Language Database Query Interfaces". Distributed at the *Workshop on Evaluation of Natural Language Processing Systems.* Wayne, Philadelphia.

[3] Bouchard, L. H., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent & P. Zweigenbaum (1992). "First Results of a French Linguistic Development Environment", *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92).* pp.1177-1181. Nantes.

[4] Bouchard, L. H., L. Emirkanian, D. Estival, C. Fay-Varnier, C. Fouqueré, G. Prigent & P. Zweigenbaum (1991). "Environnement de Génie Linguistique". Final report to the ACCT.

[5] Flickinger, D., J. Nerbonne, I.A. Sag & T. Wasow (1987). "Toward Evaluation of NLP Systems". Hewlett-Packard Laboratories, distr. at the 24th Annual Meeting of the Association for Computational Linguistics (ACL), Stanford.

[6] Grover, Claire, Ted Briscoe, John Carroll & Bran Boguraev (1989). *The Alvey Natural Language Tools Grammar (Second Release).* Tech. Report 162. Computer Laboratory, University of Cambridge.

[7] Heid, Ulrich & Elke Hildenbrand. (19??). "Some practical experience with the use of test suites for the evaluation of SYSTRAN." ???

[8] King, M. & K. Falkedal (1990). "Using Test Suites in the Evaluation of Machine Translation Systems". *Proceedings of the 13th International Conference on Computational Linguistics (COLING),* Helsinki.

[9] Klein, Judith, Ludwig Dickmann, Kader Diagne, John Nerbonne & Klaus Netter. (1992) "DiTo — Ein Diagnostikwerkzeug fuer deutsche Syntax." in: Guenther Goerz (editor) : Konvens 1992. Nuernberg 1992, pp.380–384.

[10] Klein, Judith, & Ludwig Dickmann. (1992). " DiTo — Ein Diagnostikwerkzeug fuer deutsche Syntax. Daten-Dokumentation: Verbrektion und Koordination." *DFKI Technical Report.* Saarbruecken.

[11] Nerbonne, John, Klaus Netter, Kader Diagne, Ludwig Dickmann & Judith Klein. (1993) "A Diagnostic Tool for German Syntax." in *Machine Translation.* 8. pp.85–107.

[12] Nübel, Rita. (1993). "Möglichkeiten zur Evaluierung eines kommerziellen MU-Systems: Erfahrungsbericht aus computerlinguistischer Sicht". IAI, Universität des Saarlandes.

[13] Way, A. (1991). "A Practical Developer-Oriented Evaluation of Two MT Systems", *DLL Working Papers in Language Processing*, 26, Department of Language and Linguistics, University of Essex, Colchester, UK.

APPENDIX A

Questionnaire for Test Suites

This questionnaire is not meant to be exhaustive, and on the other hand, some information may not be available for every test suite being considered.

In some sections, especially in the section on syntax, some examples (only French and English) are given. They would not necessarily constitute the best test items in a test suite, and they are by no means to be taken as exhaustively describing either the range nor the depth of the phenomena being referred to, but merely as an aid to the users of this questionnaire in interpreting the linguistic terminology with which they may not be familiar.

I. SPECIFICATION OF MATERIAL

(some information may be absent)

Name:

Author:

Institution:

Date:

Language(s) involved (source and target, if applicable):

Size (number of test items):

State of development (under development/completed):

Bibliographical References (use refer format, if available):

II. PURPOSE

Evaluation / Checking coverage

if 'Evaluation':	diagnostic/progress/adequacy
if 'progress':	specify the number of releases tested and the duration of the evaluation
if 'adequacy':	specify the requirements and test conditions

Type of system: MT, Grammar checker, ...

Type of user (tester): developer/external evaluator/future user

State of the system: under development/prototype/commercial system

III. COVERAGE

A. Breadth of coverage

For each item, specify whether a phenomenon is tested systematically or at random (S/R) and whether it is tested in depth or not (D/N).

MORPHOLOGY

inflectional:	verbal (regular/irregular)
	nominal
	adjectival (declension, comparative/superlative)
derivational:	(adverb formation/agent nouns from verbs ...)
compounding	
contraction	*aux/des* ; *doesn't/won't*
elision	*l'/qu'/d'*

SYNTAX

Specify the syntactic phenomena tested (each item as applicable to the various languages, examples are only given as a help to evaluator).

1. Sentence

declarative: positive, negative

interrogative:
- yes/no question
- WH-questions: extraction from subject/object/PP/NP complement/...
  subject: *Who came?*
  object: *Who did you see?*
  PP: *Who did you talk to? To whom did you talk?*
  NP complement: *Whose pictures did you see?*

subject inversion: (with negation, modals, some adverbs, etc...)

imperative:
- tensed
  - choice of subject, verb form
  Go! , *Let's go!*
  Allez! (understood 'you'), *Allons!* (understood 'we')
  - ordering:
  Mettez-les sur la table. Ne les mettez pas sur table.
  Allons-y! (Nous y allons)
- infinitive *Bien refermer le couvercle.*

conditional constructions:
- if-then/inversion/subjunctive/etc...
- tense and modality concordance
  If you don't go, I will.
  If you didn't go, I would.
  If you hadn't gone, I would have.
  ** If you hadn't gone, I will.*

sentential modification: adverbs
participial constructions

2. VP

agreement: with subject, with object (e.g. past participle), ...

argument structure:
- diathesis: active/passive/reflexive/middle
- subcategorisation: arguments/modifiers

syntactic structure:

- impersonal constructions:	choice of subject: null/dummy (which one: *it/there*, *il/ce*) control (interpretation of the subject of an embedded complement: arbitrary/coreferential) *Il faut partir.*, *It's hard for you to leave.*
- predicative constructions:	(NP/AP/PP predicates) control and agreement
- pronominal constructions:	reflexive/reciprocal
- pronominal complements:	linear order cliticization
- prepositional complements:	choice of preposition ordering (e.g. ditransitive alternation)
- infinitive complements:	control properties choice of complementizer
- clausal complements:	choice of mood/aspect choice of complementizer (factitive/ind.question/other) *ask [that/*0/whether/if]* ; *reply [that/0/*whether/*if]* *demander [que/si]*; *répondre [que/*si]*

verbal modification:	adverbs negation (placement, AUX-support)
verbal complex:	placement of tense/modal/aspectual auxiliaries placement of negation/adverbs/clitics

3. NP

NP specifier:

Det-N agreement
simple/complex
determiners/demonstratives/possessives/quantifiers
numerals: cardinals/ordinals
partitive construction

NP modification:

- adjective:	placement, agreement
- relatives:	restrictive/non-restrictive R: *The man, whom I had never seen before, was very angry.* NR: *The man who came to see you was very angry.* extraction from subject/object/PP/NP complement... subject: *the man who came* object: *the man I saw* PP: *the man I talked to* NP complement: *the man whose sister I talked to* temporal: *the day I came* locative: *the place I went*

NP complementation:

- PP (possession/specification/argument):	choice of preposition
- infinitive :	control properties choice of complementizer
- clausal complement (Island phenomena)	

4. AP

AP modification:	adverbs
AP complementation:	clausal and infinitive (same as VP) comparative/superlative complement (full/elliptical)

5. Coordination

coordinated structures:

S-S
VP-VP
NP-NP
AP-AP
ADVP-ADVP
Gapping
Right-Node Raising
John will buy whatever books and Mary will buy whatever records you want to get rid of.
unlike constituent coordination
John is a Democrat and proud of it.

extraction from coordinated structures

agreement with coord.struct.: determiners/Subj-Verb/reciprocals/reflexives

6. Other constructions

Dislocation	*This guy, I hate him.*
Topicalisation	*This guy, I hate.*
Cleft (IT-cleft)	*It's ice-cream that Mary likes.*
Pseudo-cleft (WH-cleft)	*What Mary likes is ice-cream.*
Extraposition	*Its' amazing, the amount of money he'll spend!*
Parentheticals	(comma-bounded, dash-bounded)
Direct Speech	

7. Syntax/Semantics

Anaphor Resolution: syntactic restrictions on coreference:
- reflexives: *Mary told Bill about him/her/herself/*himself.*
- c-command: *Mary[i] said she[i] would leave.*
** She[i] said that Mary would leave[i].*

VP ellipsis/sluicing/gapping/pseudo-gapping

modality, tense and aspect: concordancy in main and subordinate clauses
He asked whether I would go/ I was going.
He said he will go.
We can go to the movies when you are finished.
We will go to the movies when you are finished.
** We will go to the movies when you will be finished.*

cooccurrence with adverbs
He asked whether I would go tomorrow.
** If you don't go yesterday, I will.*

Negative Polarity Items (negative concordancy)
Does anybody know the answer?
Does somebody know the answer?
Nobody knows the/any answer.
Somebody doesn't know the/any answer.
** Anybody doesn't know the/any answer.*
** Nobody doesn't know the/any answer.*
Somebody knows the answer.
** Somebody knows any answer.*

Ambiguity:	
- negation:	scope (sentential/verbal/nominal) *He reads no books.* *He doesn't read books.* *He doesn't read.*
- PP attachement:	NP/VP/S *He saw the man with a telescope.* *He saw the man by the window.*
- scope of adverb:	sentential/verbal modification *He told me he would get the trash out yesterday.*
- quantifier scope:	*Every student reads a book.*

8. Idiomatic expressions (not compounds)

En. *to kick the bucket, to keep tabs on*

Fr. *casser sa pipe, prendre une veste*

SEMANTICS

(give other semantic distinctions which may be drawn)

verbal polysemy inter-sentential anaphora

EXTRA-GRAMMATICAL

1. punctuation: full stop, comma, inverted comma, colon, dash, etc...
2. formatting: titles, headings, lists, quotations, insertions, etc...
3. terminography: acronyms, abbreviations, proper names, place names, titles measure phrases, formulae, foreign language words, etc...
4. robustness: behavior with unknown words or characters

Is anything else tested?

B. Depth of coverage:

ill-formed data:

- tested systematically/randomly/not at all
- what kind of ill-formed data (spelling errors, punctuation errors, missing words, etc)
- indicate whether the impact of ill-formed data on the results has been measured and interpreted (and if so, is this available?)

interaction and co-occurrence of phenomena

- describe whether there is a system in testing cooccurrence,
- how many cooccurring phenomena occur in test data, etc.

level of detail
- can you specify how fine-grained the test data are?

exhaustivity
- can you specify in what sense test data are incomplete?

IV. PRESENTATION (logical and physical)

Are the test data annotated?
- in what way?
- to what degree?
- are the annotations systematic?
- can they be reused?
- give a list of the annotations

Are the test data documented?
- describe documentation

Are the test data divided in a logical user-friendly way (not merely listed)?
- what is the schema used?

V. CHARACTERISTICS OF THE SOURCE

What source are the test data drawn from?

Are the test data drawn up with a specific application in mind? which one?

Type of vocabulary used for the test data:
- subject area
- size

VI. OTHER

Is the test suite design dependent on a linguistic theory?
- which theory?
- to what extent is it portable to another linguistic theory?

Is the test suite design dependent on a theory of testing and evaluation? - which theory?

Is the test suite supported by any tools?
- database
- construction tools

- other

Is the test suite flexible? (i.e. is it easy to replace test data with equivalent but different data?)
What is needed to change the test suite (besides an editor...) ?

Has the test suite been applied?

- Are the results of using it available? where?

- Are interpretation of the results available or documented? where?