The Coordinator's Final Report on the First Morpholympics

March 7 and 8, 1994
The University of Erlangen-Nürnberg

Roland Hausser
Gesellschaft für Linguistische Datenverarbeitung *GLDV*
Working Group *Parsing in Morphologie und Syntax*

Most scientific conferences are exciting and unpredictable events for active participants and organizers alike. This proved to be especially true of the 1. Morpholympics, a new type of conference combining the presentation of scientific papers with a competition for testing software on test samples. Being the first event of its kind, participants, jury, and organizers all moved in uncharted waters.

This report analyzes some of the experiences gained during preparation and realization of the 1. Morpholympics. Section 1 discusses different methods of evaluation. Section 2 proposes a standardized procedure for preparing testing devices. Section 3 explains where difficulties with testing devices may arise. Section 4 discusses procedural issues. Section 5 summarizes the experience with a proposal for standardizing procedures in future Morpholympics.

1 Preparations

1.1 Standardized Presentations

To maximize consensus among potential participants at the 1. Morpholympics, a preparatory meeting was held on October 14 and 15, 1993, in Erlangen. It was attended by researchers from 12 universities in Austria, Germany, and Switzerland, all experienced in computational morphology. After the on-line presentation of 10 systems,[1] the group discussed and reached agreement on which aspects of morphological parsers should be evaluated.

The results were published as part 2 of the announcement of the 1. Morpholympics.[2] This *Questionnaire for the standardized presentation of systems* presents detailed questions on the following points:

2.1 *Conceptual criteria*
2.1.1 Declarative specification of lexical entries and rules
2.1.2 Relation between lexical entries and word forms
2.1.3 Transparency and linguistic motivation of the rules
2.1.4 Morpho-syntactic analysis (categories)
2.1.5 The handling of generation
2.1.6 Application to other natural languages

2.2 *Technical design and practical use*
2.2.1 Conceptual goal of the design
2.2.2 Portability of software and data
2.2.3 Interfaces to syntax and semantics
2.2.4 Aiding the user
2.2.5 Limits on the size of the system
2.2.6 Interface to non-ASCII characters
2.2.7 User friendliness of the 'turn around'
2.2.8 The state of the documentation
2.2.9 Availability and maintenance

It was assumed that the information provided on each of these points in combination with the measurements on data coverage and speed would be sufficient to provide a solid, empirical basis for an objective evaluation of the participating systems.

[1] See Seewald 1993, LDV-Forum 10.6, p. 6 - 16.

[2] The German version appeared in LDV-Forum, Vol 10.6 (December 1993), p. 17 - 23, the English version was made public on various electronic lists and is now available via anonymous ftp from sol.linguistik.uni-erlangen.de.

1.2 Approaches to evaluation

Right before the actual contest, members of the jury expressed the feeling that the criteria for evaluating different systems had not been defined clearly enough. Instead of further elaborating, weighting and formalizing the criteria established during the preparatory meeting, the jury decided on a more spontaneous approach, selecting and balancing what they felt to be the most impressive.

This is a possible approach, corresponding to traditional practice. For the longterm success of a competition like the Morpholympics, however, it is important that participants and jury alike operate with a common and widely accepted set of clearly defined assumptions and expectations.

The first and most basic step in this direction is bringing the test data into a canonical form. The second step is to evaluate all competing systems systematically and equally with respect to their performance in each of the tests. The third is to arrive at a decision based on steps one and two.

2 Canonical form of testing devices

2.1 Types of tests

In order to evaluate morphological parsers with respect to their *completeness 01 coverage, speed, quality 01 analysis,* and *quality 01 implementation* they should be tested using the following five types of testing devices:

1. text of written language (w-text)
2. text of transcribed spoken language (s text)
3. list of word forms derived from texts (t-list)
4. questions for evaluating quality of analysis (a-questions)
5. questions for evaluating quality of implementation (i-questions)

These different types of testing devices make different demands and test for different properties.

Analysis of the w-text requires that the system can handle the structure characteristic of written texts, such as headings, punctuation, line breaks, end-of-the-line hyphenation, and special characters. For the analysis of the s- text, the system must handle the conventions for transcribing spoken language.

The t-list contains each word form only one e. For this reason coverage can be tested for a much larger number of forms than in a text of comparable size. Also, because of the uniqueness of forms, the use of a *cache* for high frequency words or for reusing previously analyzed forms has little or no advantage in the case of t-lists. Therefore, a system's speed will usually be considerably slower on word lists than on a comparable text. t-lists are to be generated automatically from a corpus of texts, possibly using certain structural or statistical principles of selection.

The set of a-questions should probe the quality of linguistic analyses by presenting word forms which relate to the topics presented in the questionnaire under 2.1 *Conceptual Criteria.* Each a-question should be furnished with comments clarifying the purpose of parsing the test forms.

The set of i-questions, finally, should address the topics presented in the questionnaire under 2.2 *Technical design and practical use.* They may be extended into a subtest in which participants develop a grammar for a small, clearly defined set of morphological data from a little known language. By publicly demonstrating the adaptation of each system to the same set of new data, the systems' practical handling, the nature of their rules, and their conceptual approach to morphological analysis may be demonstrated most clearly.

2.2 Preparation of Testing Devices

During the preparatory phase of the 1. Morpholympics, it seemed sufficient to simply use well selected pieces of text or lists of word forms as test samples. After all, the systems should perform in as natural a situation as their normal practical use would require. While this reasoning is generally true, it overlooks that the *comparison* of different systems is difficult, if not impossible, unless the samples have been carefully prepared for the purpose of testing.

The metamorphosis of a piece of nondescript on-line data into a testing device is brought about by embedding the data into a standard structure. This structure consists of a two part header at the beginning of the test file[3] and markers indicating beginning and end of different kinds of data. By stating the header and the begin/end-markers within the formal convention of SGML comments,[4] the prepared test file may serve as input to a morphological parser without any need for further editing,[5] provided the morphological parser knows how to handle the SGML comment convention.

The data embedded into the structure of a test file are not to be modified in any way. The structure of a test sample depends on its type. We begin with a detailed description of the structure of w-text test samples, to be followed by briefer descriptions of texts and t-lists and a-questions.

2.2.1 Preparation of w-texts

A textual testing sample prepared from a written text consists of part 1 of the header - providing information on the origin of the sample, part 2 of the header - providing essential statistics, part 1 of the data, consisting of the list of ill-formed word forms found in the text and part 2 of the data, consisting of the text itself. The two data parts are clearly set off by begin/end-markers consisting of declarations beginning with upper case letters of the Latin alphabet.

The declarations of the header are filled out by the researcher preparing the sample.

```
<!- - 1. Type of test sample: w-text - ->
<!- - 2. Name and address of researcher selecting
the sample: - - >
<!- - 3. Time, place and occasion of creating the
test sample: - - >
<!- - 4. Reason(s) for selecting text as test sample: -->
<!- - 5. Origin of text: (author, date, publisher)
->
<!- - 6. Structure of the text: (e.g. SGML) - ->
<!- - 7. Coding used for special characters: (e.g.
ASCII) - - >

<!- - a. Method for counting word forms: - -> <!- -
b. Total number of word forms: - ->
<!- - c. Number of well-formed word forms: - ->
<!- - d. Number of ill-formed word forms: - -> <!-
- e. Percentage of well-formed word forms: ->
<!- - A. Begin list of ill-formed word forms - - >

        form1
        form2
        form3


<!- - A. End list of ill-formed word forms - ->
<!- - B. Begin w-text - ->
W-TEXT
<!- - B. End w-text - ->
```

The general information about the origin of the sample begins with a declaration specifying the '1. Type of the sample.

By giving his or her '2. Name and address' a researcher takes charge as the author of the testing device. Stating the '3. Time and place' of creating the test sample and the '4. Reasons for selecting' a given text are useful for developing a taxonomy of different types of test samples.

An exact specification of the '5. Origin of text' is necessary for future comparisons with other test samples. Questions regarding the exact nature of misprints, hyphenations, typographical and dialectal idiosyncrasies and other properties important for the performance evaluation of morphological parsers can only be resolved by being able to go back to the specified origin of the document.

An explicit specification of the conventions used for representing '6. Text structure' and '7. Special characters' tells a user right away whether a given system is able to handle the sample. This information may also be used directly by morphological parsers capable of interpreting SGML declarations, including SGML comments in the position of the header .[6]

[3] Addition of the header information in the test file itself, rather than a separate Readme file, is also practical for later references, when only certain samples from a test set may be selected for new test runs, discussion or comparison.

[4] See Herwijnen *1990,* p. 72.

[5] E.g. removal of the header.

[6] For example, there are currently at least five different conventions for representing Umlaut in German, according to which, e.g., the preposition

The second part of the header provides numerical information on the number of well-formed and ill-formed word forms contained in the sample. It begins by stating the 'a. Method of counting word forms.,7 If no official method of word count is stated, the numbers for one and the same text arrived at by different systems may vary by as much as 18% (see section 3.1).

Giving the official 'b. Total number of word forms' allows competing morphological parsers to calibrate their word count algorithms. The declaration 'co Number of well-formed word forms' provides the mark morphological parsers should achieve when analyzing the sample. The declaration 'd. Number of ill-formed word forms' is redundant in light of band c, and therefore suited to indicate whether the numbers are consistent.

Stating the 'e. Percentage of well-formed forms' gives a handy guide line for the initial evaluation of coverage by a morphological parser. This is because most modern parsers automatically provide statistical information at the end of an analysis, including the *percentage of analyzed word forms.* The percentage of analyzed word forms provided by the parser should equal the official percentage of well-formed forms (and not 100% of the grand total of word forms).

The two part header is followed by two kinds of data. The first consists of an explicit official list of the ill-formed word forms contained in the text. This list is the basis for the numerical information in declarations d and e. By treating the list as part of the sample data, it is analyzed during parsing.

Finding the candidates for the list of illformed word forms in a larger text is easy enough by using a suitable morphological parser. The final decision on whether a word form is ill- formed or not may sometimes turn out to be a matter of different opinions, however. In this case, the form should be added to the list, followed by a comment consisting of a "?' and the standard variant of the form.

The second kind of data consists of the written text itself. The beginning and the end of the two kinds of data are clearly marked by SGML comment declarations. This helps in the interpretation of the testing device. Marking the end of the text is also useful for checking whether the test sample is complete when transmitted over the net.

2.2.2 Preparation of s-texts

The preparation of s-tests closely resembles that of w-texts. Differences arise only in
the lines 1 and 6 of the header and the begin/end-markers of the text data:

<!- - 1. Type of test sample: s-text - -> ...
<!- - 6. Method/standard of transcription: (e.g., CHAT) - ->

<!- - B. Begin s-text - ->
S- TEXT
<!- - B. End s-text - ->

The main difference between w-texts and s-texts is that the textual structure of w-texts is provided by the author and/or publisher, whereas the structure of s-texts is imposed during the transcription.

2.2.3 Preparation of t-lists

The purpose of t-lists is to check the coverage of morphological parsers on a large set of word forms automatically derived from a corpus of texts representing a certain domain. The structure of t-lists should be such that each form to be analyzed occurs only once and is written into a separate line.

Like text samples, t-lists consist of a two part header and two kinds of data. The first kind of the data consists of a list of illformed word forms, the second of a list of

für may occur as 'für', 'f\374r', 'für', 'f' 'ur', and 'f}r'. Reading through the header of the text to be analyzed, a suitably extended system may determine and activate the specific convention required by the text. Capabilities like this should be added to the set of i-questions (2.2.5) in future Morpholympics.

[1] A widely available method of counting the number of word forms in an on-line text automatically is wc of Unix. It should be noted, however, that wc is in many ways rather empty. For example, wc counts the parts of hyphenated words separately also, punctuation signs surrounded by spaces, e.g. '-' are counted as additional word forms, whereas those following a word without a space, e.g. '... sample. ' are not. Instead of wc a simple, linguistically motivated algorithm should be used, which disregards punctuation signs and treats the parts of end-of-the-line hyphenation as one word form.

well-formed word forms. To facilitate orientation by the user, the two sub-lists should each be structured according to the same ordering principle, such as alphabetical order. t-lists differ from text samples in that both kinds of the data are lists which are moreover disjoint.

<!- - 1. Type of test sample: t-list - ->
<!- - 2. Name and address of researcher selecting the sample: - - >
<!- - 3. Time, place and occasion of creating the test sample: - ->
<!- - 4. Reason(s) for selecting list as test sample: -->
<!- - 5. Origin of text or corpus from which list was derived: - - >
<!- - 6. Structure of the list: (e.g. alphabetical order) - - >
<!- - 7. Coding used for special characters: (e.g. ASCII) - - >
<!- - 8. Method by which list was derived from corpus: - - >

<!- - a. Total number of word forms: - ->
<!- - b. Number of well-formed word forms: - ->
<!- - c. N umber of ill-formed word forms: - - >
<!- - d. Percentage of well-formed word forms: ->
<!- - A. Begin list of ill-formed word forms - - >

Form1
form2
form3

<!- - A. End list of ill-formed word forms - - >
<!- - B. Begin list of well-formed word forms - ->

Form
1
form2
form3

<!- - B. End list of well-formed word forms - ->

The first part of at-list header differs from that of w-text and s-text headers in lines 1, 5 and 6, which are straightforward adoptions to the different sample type. Of special interest is the additional declaration in line 8, which specifies the method by means of which the t-list was constructed from a given set of texts.

For example, 'listl', which served as the t-list at the 1. Morpholympics, was constructed automatically as a sub set of word forms of the LIMAS corpus. The subset was formed by selecting only word forms from the open classes with a frequency of eight or more occurrences in the corpus.

The second part of at-list header does not state the method of counting word forms because this task is trivial, given that lists write each form into a separate line, contain no punctuations signs and do not use end-of-the-line hyphenation. The declarations a, b, c and d correspond to b,c,d, and e in text samples. Because word forms are unique in t-lists, the numbers of (a) word forms, (b) well-formed word forms, and (c) ill-formed word forms have a different status as compared to texts.

2.2.4 Preparation of a-questions

The sample types w-text, s-text and t-list have in common that they contain data or are based on data - that were produced and made public solely for regular purposes of normal communication. Because alterations of these data are not permitted for reasons of scientific method, the researcher has no influence on the specific word forms contained in such a test sample, apart from choosing a particular text or corpus and, in the case of t-lists, a particular method of automatic selection.

In contrast, a-questions are a testing device hand-made by linguists to check a parser's handling of specific phenomena in the morphology and orthography of a natural language. This freedom of chasing whatever forms the author of the list finds interesting should be complemented by comments which clarify the testing purpose of each form.

In order to allow the author of questions to present her/his data in the most transparent and perspicuous manner, the data are loosely organized into a list of 'data items'. Each data item consists of a statement describing its purpose and an open list of word forms. The word forms to be parsed are embedded into a format that can be handled by the systems.

Like the t-list, the a-questions are preceded by a standard header .

<!- - 1. Type of testing device: a-questions - -> <!- - 2. Name and address of researcher selecting the sample: - - >
<!- - 3. Time, place and occasion of creating test device: - - >
<!- - 4. Coding used for special characters: (e.g. ASCII) - - >

<!- - a. Total number of word forms: - ->
<!- - b. Number of well-formed word forms: - ->

```
<!- - c. Number of ill-formed word forms: --> <!-
- d. Percentage of well-formed word forms: ->


<!- - A. Begin list of data items - ->
        <!- - BI. Begin data item - -> <!-
        - Purpose of data item: - ->
                form!
                form2 <!- - * - ->
                form3
                form4 <!- - ? - ->
        <!- - BI. End data item - ->

        <!- - B2. Begin data item - ->
        <!- - Purpose of data item: - ->
                form! <!- - * - ->
                form2<!- - ? - ->
                form3 .
                form4
        <!- - B2. End data item - ->


<!- - A. End list of data items - ->
```

Compared to t-lists, part 1 of the header differs in lines 4 and up. This is because a-questions are hand-made. Thus there is neither a text from which they derive, nor any method of derivation.

Each item of a-questions should be interpretable as checking a specific aspect of quality of linguistic analysis. For example,

> < - - Purpose of data item: The following examples are intended to show the handling of valency frames in verbs: - - >

or

> < - - Purpose of data item: The following examples are intended to show the handling of 'Fugenelemente' in the composition of nouns: - ->

The word lists following the statement of purpose may present well-formed and illformed word forms in any order the author finds suitable to her/his purpose. Because a-questions do not present the ill-formed word forms as a separate list - in contrast to w-texts, s-texts and t-lists -, it is imperative that ill-formed word forms are clearly marked by a comment. Based on the number of such comments, the number of ill-formed word forms can be determined automatically.

In explaining the purpose of an aquestion, the author may have to make explicit which grammaticality judgement or morphosyntactic characterization of a word form is assumed. If such assumptions spark controversies, it is certainly better than leaving the testing purpose of the data unexplained. The a-questions should help guiding and promoting the discussion of central topics in the community of morphological parsing.

2.2.5 Preparation of i-questions

The final test data used at the 1. Morpholympies, called textl, text2, listl, and list 2, corresponded more or less to the concepts of a w-text, an s-text, at-list and a set of a-questions, respectively. The concept of a set of i-questions, on the other hand, had not yet evolved at the time. For this reason, no specific experiences were made at the 1. Morpholympics that would guide in the formulation of standards for i-questions.

In light of the general experience, however, it seems important to test systems for automatic word form recognition with respect to their ability to adapt to new data. This may be done by presenting a small, clearly defined set of morphological data from a little known language in the set of i-questions. The description of the data should be followed by a list of well-formed and ill-formed word forms.

The task of each participating system is to write a grammar for these data and demonstrate its adequacy by parsing them. In this context systems may be evaluated with respect to additional theoretical and practical issues like the following:

- How well does the system separate between the rules used for a specific application and the general parser applying these rules?
- Does the system use a declarative rule format and how readable is its grammar for the set of new data?
- Does the system generate its own error messages or does it rely solely on the debugger of the programming language used?

. How easily does the system adapt to the handling of new special symbols?

. How long does it take to perform all the tasks requested by the i-questions?

The procedure of extending a system to handle the set of i-questions may be organized as a public performance, where the representative(s) of the system explain(s) each step, elaborating the linguistic and technical motives behind it.

3 Testing problems

As final test data for the 1. Morpholympics, the jury provided four files, called textl (2375 word forms[8]), text2 (1674 word forms), listl (3817 word forms) and list2 (282 word forms). While these files[9] corresponded more or less to the concepts of a w-text, an s-text, at-list and a set of aquestions, respectively, they did not support the standards described 2.2.1 - 2.2.4. This created the following problems for evaluating the analyses of the participating systems:

3.1 Official word count totals

The final test samples did not provide numbers of word forms for each of the four test samples and no method for arriving at such numbers was agreed upon. Consequently, the numbers for the total word forms given by the eight participants varied widely. For example, for text! the eight participating systems submitted the following word counts to the judges: 2023, 2082, 2142, 2156, 2375, 2380, 2400.

The difference between the lowest count of 2023 and the highest of 2400 is a 377 word forms and amounts to a whopping 18.5%. This must be seen in light of the fact that the difference between the best percentage (95.5%) and the worst percentage (86.0%) of 'word forms analyzed' is only 9.5%. Thus the range of difference between different systems regarding the word count is twice as large as the difference in their respective percentages of analyzed word forms.

3.2 Official numbers of ill-formed word forms

The final test samples did not specify how many of the word forms in a sample happen to be ill-formed and should therefore not be recognized. This affected the proper evaluation of the percentage of recognized word forms.

For example, a quick examination of textl found the following 23 occurrences of misprints and uncommon abbreviations.

Pfarr, fur, Schriftum, WV (13)[10], syntematischen, Artiekls, Mei-nungeen, Millioen, gefdhreden, staatslichkirchlicher , seitigen

This amounts to 1 % of the actual 2020 word forms of textL Also of interest in this connection is an additional 1 % of unusual names and foreign words.

In list2 with its total of 280 actual word forms the following 15 problem forms can be found:

molket, geretten, gesagen, geruft, vorbeigerannene, genennt, ankoemmt, voreingenommt, lieferen, schlotzen, schlotzte, geschlotzt, %anrufe, %anrufend, % angerufen[11]

This amounts to 5.3% of the word form total of list2.

3.3 Official identification ill-formed ofword forms

ill-formed word forms were neither marked in the test samples nor presented as a separate list as part of the data. Regarding list2, for example, it remains a mystery whether forms like *molket, ankoemmt* or

[8] These word counts are based on wc. Note, however, that text! had been edited so that the punctuation signs ' . , ; : "_, appear between two spaces, thus being counted as separate words by wc. Therefore, the real number of word forms of text! is only 2020.

[9] The test data used at the 1. Morpholympics are available via anonymous ftp from so@linguistik.uni-erlangen.de in the directory 'morpholympics. '

l OWV may not be found in the *Wahrig Deutsches Wörterbuch* and is yet to be deciphered.

[11] Apparently, three lines were intended to be commented out, which did not stop most parsers.

schlotzte were intended to be well formed or not. Consequently, it is unknown whether their recognition and analysis by a morphological parser should be counted as an achievement or as a mistake.

3.4 Evaluating the parsing of large texts or lists

It turned out that some systems used three pages and more for the analysis of a single word form. Consequently, files containing the analyses of text 1, text2 and list 1 were often huge and their perusal by the jury was frustrating due to a wealth of low level information and a concomitant lack of structure. Obtaining objective results on a system's degree of coverage by browsing through the associated files turned out to be impossible in the short time available.

Nevertheless, the parsing of large test files can serve as a simple, fast, and precise instrument for determining the quality of coverage by any number of competing systems if the following conditions are met:

- The test samples have been properly prepared, providing standardized word counts, explicit lists of ill-formed word forms, and correct official percentage numbers of well-formed word forms.
- The competing systems use the same method of word count and automatically provide the statistics described in 4.3.12

The list of ill-formed word forms provided at the beginning of w-texts, s-texts and tlists will show at a glance whether a system accepts ill-formed input or not.

3.5 Statements of purpose in a questions

As the only sample of the final test data constructed by hand, list2 most dozily resembled a set of a-questions. Unfortunately, however, there were no statements indicating what the word forms in list2 were being tested for. Many possibilities exist: Specification of valency structure? Coding of separable prefixes? Formation of past participle with or without ge-? Classification of certain forms as both indicative and subjunctive?

By. its very nature, the interpretation of a set of a-questions goes far beyond the counting of recognized word forms and should help guide the discussion on issues of linguistic theory and representation. Without a dear statement of purpose, however, it is virtually impossible to evaluate the analyses of word forms in an a-question, as produced by the different systems.

3.6 Balance of phenomena tested in a-questions

List2 contained only inflectional forms of verbs. A thorough evaluation of different morphological parsers with respect to their quality of analysis should be based on a well- balanced and linguistically motivated check list which represents different phenomena from the areas of inflection, derivation and compounding. Furthermore, the inflectional data should take into account nouns, verbs and adjectives/adverbials.

Given the limited time for deliberation and the unwieldy nature of the analysis files for w-text, s-text, and t-lists it is unlikely that a jury can evaluate the quality of word form analyses of different systems simply by browsing through the respective analysis files. Instead, the quality of word form analysis should be based on a carefully prepared set of a-questions.

3.7 Remark

It would have been easy enough to bring the test data used at the 1. Morpholympics into the canonical forms specified in 2.2.1 - 2.2.4, thus avoiding the difficulties of evaluation described above. Also, participants could have easily adapted to a common method of counting word forms. Unfortunately, however, no standards for the format of test samples and their evaluation had yet been written at the time.

[12] For reasons of reliability and speed one may want to program a procedure for automatically ranking different systems with respect to their coverage of samples in canonical form.

4 Questions of procedure

4.1 Mode of participation

Participation at the 1. Morpholympics was open to any person or team that followed the rules of registration, installed the system in question via remote login, signed a publication agreement and turned in a standard questionnaire describing various theoretical and practical aspects of the system. This liberal procedure proved to be successful in that it resulted in a lively group of professionals from some very different necks of the woods.

4.2 Advancing distribution of written presentations

All the systems presented were of good quality and raised a wealth of interesting issues. Unfortunately, however, because of a strict time table and because most of the systems were not previously known to the jury as well as the other participants, there was not sufficient opportunity to attain a deeper understanding.

To improve on this state of affairs, the written presentations should be sent to the coordinator four weeks before a Morpholympics (rather than being turned in during the preparatory meeting at the beginning of the conference). The coordinator collects these presentations into a volume and sees to it that duplicates of this volume be sent to each member of the jury.13 Other persons may also obtain copies of this volume, upon request and for a fee to cover copying and postage.

4.3 Supporting standards and automating statistics

The questionnaire requires participating systems to parse a set of preliminary test data and to append the results to the questionnaire. At future Morpholympics, these test runs should be used to calibrate systems to a common method of computing word form totals for arbitrary test samples of canonical form.

Furthermore, it is to be made obligatory for all systems participating at a Morpholympics that they *automatically* provide the following measurements at the end of an analysis file:

Total number of word forms encountered in the sample
N umber of word forms successfully analyzed
Number of word forms not recognized
Percentage of analyzed word forms
N umber of analyzed word forms per second

It is not difficult to add this feature to systems which do not yet have it. Automatic statistics are more reliable and quicker than calculations by hand, especially in the hectic atmosphere of a competition. Also, they will prove to be quite useful for practical work outside the context of the Morpholympics.

4.4 Only one parse per sample on newly loaded system

The measurements of an official parsing test must be based on a newly loaded system, parsing the test files in a given order and using exactly *one* test run per sample. There are at least two scenarios where using data from a second parse of a given sample leads to an improper manipulation of test results.

The first, which happily was not encountered during the 1. Morpholympics, has general technical reasons: Running a system on a sample for the first time requires the reading of data from the disco When parsing the same data a second time without reloading the system, the information read from the disc during the first analysis will still be available in the run time memory and thus result in a considerably faster timing.14 The measurements of the second parse are not relevant because in practical applications the user will not run the system twice just to enjoy a seemingly faster parse.

[13] An even simpler method for the coordinator is to install the dvi-files of the presentations in an ftp directory from which the judges can obtain the presentations electronically and print them out at their respective offices.

[l4] In the order of 4,000 versus 10,000 word forms/second in the case of the LA-Morph system.

The second scenario is more specific in that it depends on a system-dependent distinction between two separate phases in the parsing of a word form, called 'recognition' and 'analysis'. The first phase consists in a quick check whether the word form at hand is recognizable at all. If it is, a full analysis may be computed in the second phase.

Such a system may be run using only the first phase. In this case, a parse is about 5 times faster than when word forms are really analyzed. If the restriction to exactly one parse per sample is not enforced effectively, the participant running the test might be tempted to compose his results from the best of two different parses and then forget to mention that the timings are not those of the analyses.15

4.5 Variation of test data

In order to give potential participants an idea of what kind of test samples to expect during the competition, a set of preliminary samples was made available two months before the Morpholympics.16 These preliminary samples generated some discussion regarding the conventions of text structure, the coding of special symbols, and the handling of end-of-the-line hyphenation. To get the Morpholympics off to a gentle start, the final test samples were taken from the same domain as the preliminary samples and even edited to regularize and simplify various aspects of coding.

As a consequence, the participating systems varied less than 10% of the total number of word forms in their coverage of text samples. At future competitions, the range of domains should be extended so that a high degree of coverage is more difficult to attain. Furthermore, the systems' handling of different coding conventions should be tested.

In fairness to other participants, care must be taken that the nature of the final test samples is not leaked before the competition. This holds in particular with respect to the t-list, which may be reconstructed from description.

4.6 Testing portability

During the preparatory meeting in October 1993 it was agreed that systems should be tested on three different platforms to demonstrate port ability and to compare measurements. The platforms mentioned were a workstation, a Macintosh, and a PC. Regarding the operating systems nothing specific was said. It was assumed, however, that the operating systems usually associated with these respective types of hardware would be used.

This lack of specificity in the announcement was inconsequential with respect to the choice between DOS and WINDOWS on the PC. But by the time of the l. Morpholympics, the recently introduced LINUX operating system had been discovered to allow easy transfer of programs developed on Unix work stations to the PC, requiring no major adaptations and running at a quite respectable speed.

Even though the running on three different platforms had been announced as an obligatory part of the competition, 5 systems ran exclusively on Unix work stations and 1 system ran exclusively on the PC under DOS. Of the remaining two systems, one tested on the workstation, the Macintosh, and the PC under LINUX, while the other ran an additional fourth test, namely on the PC under DOS.

The practice of routinely adapting a piece of software to run under varying conditions is of great theoretical and practical benefit. It should therefore be encouraged in future Morpholympics, at least in the form of adding or subtracting points in the final evaluation of a system. That the timing of systems on different platforms is meaningful and interesting is demonstrated by the fact that the systems taking first and second place on the workstation under Unix were reversed on the PC under LINUX.17

[15] Helping to ensure that the results of test runs need not be questioned is another reason why the various measurements (see 4.3) must be calculated automatically.

[16] See 1.1.6 of 'Organization and implementation' in the announcement of the 1. Morpholympics.

[17] See Coordinator's announcement of results, available via anonymous ftp from sol.linguistik.uni-erlangen.de.

5 Summary

To ensure that the evaluation of different systems is based on objective criteria, standardized procedures for the quantitative and qualitative measurement of performance should be advanced. The basis for such procedures is the preparation of different testing devices, consisting of three sets of data to be parsed and two sets of questions for checking specific aspects of quality.

Regarding quantitative measurements of software performance, it has been argued that small differences in speed, measured as the number of word forms per second, and coverage, measured as the percentage of word forms recognized, are not really meaningful. The same could be said about sporting events like the 100 meter sprint. From a practical viewpoint of daily life, a difference of a fraction of a second is indeed not really significant, but for winning the competition it is. The important property of such small differences is that it is (1) unquestionably objective and (2) agreed on by all participants. Runners of all shapes and sizes with different views on life and morals can adapt to and focus on this one parameter.

In the case of the Morpholympics, the evaluation is somewhat more complicated and more balanced because there are altogether four parameters relative to which systems are measured. These are (1) coverage, (2) speed, (3) quality of linguistic analysis, and (4) quality of implementation. The first two are measured quantitatively based on parsing the w-text, the s-text and the t-list. The latter two are evaluated qualitatively by using the a-questions and the i-questions, respectively.

Like the test samples, the catalogues of questions should be written by the jury in advance and kept under lock and key till the day of the contest. Just as all competing systems are tested and ranked with respect to coverage and speed on a given set of test samples, all systems should be tested and ranked with respect to explicit sets of questions checking for quality of analysis and implementation.

The preparation of test data add catalogues of questions prior to a Morpholympics requires a certain amount of work from the jury. However, given a jury of 5 judges, each would have to prepare only one of the 5 testing devices.

The benefits resulting would be great. Apart from an evaluation procedure beyond reproach, there would be precise measurements suitable to serve as bench marks for future systems. Even more importantly, the explicit questions for measuring quality of analysis and of implementation will direct attention to concrete problems of linguistic description and set standards for empirical analysis in future research.