Chapter 13

Evaluation

13.1 Overview of Evaluation in Speech and Natural Language Processing

Lynette Hirschman[a] & Henry S. Thompson[b]

[a] MITRE Corporation, Bedford, Massachusetts, USA
[b] University of Edinburgh, Scotland

Evaluation plays a crucial role in speech and natural language processing, both for system developers and for technology users. In this section we will introduce the terminology of evaluation for speech and natural language processing and provide a brief survey of areas where it has proved particularly useful, before passing on to more detailed case studies in the subsequent sections.

13.1.1 Introduction to Evaluation Terminology and Use

We can broadly distinguish three kinds of evaluation, appropriate to three different goals.

1. **Adequacy Evaluation**
   This is determination of the fitness of a system for a purpose—will it do what is required, how well, at what cost, etc. Typically for a prospective user, it may be comparative or not, and may require considerable work to identify a user's needs. One model is consumer organizations which publish the results of tests on, e.g., cars or appliances, and identify *best buys* for certain price-performance targets. This also goes by the names *evaluation* and *evaluation proper*.

2. **Diagnostic Evaluation**
   This is production of a system performance profile with respect to some taxonimization of the space of possible inputs. It is typically used by system developers, but sometimes offered to end-users as well. It usually requires the construction of a large and hopefully representative *test suite*. It also goes by the name *diagnosis*, or by the software engineering term *regression testing* when used to compare two generations of the same system.

3. **Performance Evaluation**
   This is measurement of system performance in one or more specific areas. It is typically used to compare like with like, whether two alternative implementations of a technology, or successive generations of the same implementation. It is typically created for system developers and/or R&D programme managers. When considering methodology for measurement in a given area, a distinction is often made between *criterion*, *measure* and *method* (see below). It also goes by the names *assessment*, *progress evaluation*, *summative evaluation* or *technology evaluation*.

When systems have a number of identifiable components associated with stages in the processing they perform, it is important to be clear as to whether we approach the system as a whole, or try to evaluate each component independently. When considering individual components, a further distinction between *intrinsic* and *extrinsic* evaluation must be respected—do we look at how a particular component works in its own terms (intrinsic) or how it contributes to the overall performance of the system (extrinsic). At the whole system level, this distinction approximates to the performance evaluation/adequacy evaluation one, where intrinsic is to extrinsic as performance evaluation is to adequacy evaluation.

A distinction is often drawn between so-called *glass box* and *black box* evaluation, which sometimes appears to differentiate between component-wise versus whole-system evaluation, and sometimes to a less clear-cut difference between a qualitative/descriptive approach (*How* does it do what it does) and a quantitative/analytic approach (How *well* does it do what it does).

Adequacy Evaluation

As speech and natural language processing systems move out of the laboratory and into the market, it is becoming increasingly important to address the legitimate needs of potential users in determining whether any of the products on offer in a given application domain are adequate for their particular task, and if so, whether any of them are obviously more suited than the others. If we reflect on the way similar tasks are approached in other fields, we observe what we can call the Consumer Reports paradigm, which does not necessarily aim at actually *identifying* the *best* system, but rather at providing comparative information which allows the user to make an informed choice. Techniques from both diagnostic and performance evaluation may be called on to achieve this aim, but are unlikely to be sufficient in themselves—for example, assessing customisability may be of fundamental importance in determining adequacy to a particular user's needs, but is unlikely to be addressed by existing diagnostic or performance evaluation methodologies.

The term *formative evaluation* is used in the field of human-computer interaction to refer to a collection of evaluation methodologies more closely related to both adequacy evaluation and to diagnostic evaluation in our terms. The goal of formative evaluation is to provide diagnostic information about where a given system succeeds or needs improvement, relative to its intended users and use. The role of formative evaluation is to influence and guide system design, as opposed to performance evaluation or summative evaluation, which rates systems relative to each other, or relative to some *gold standard* such as human performance. During system development, user trials of system prototypes or alternative assessments of user interface functionality are conducted, in which more or less formal measurements of usability are recorded (e.g., via study and measurement of user actions performing some representative set of tasks, possibly coupled with interviews). We see considerable potential for importing some of these techniques into adequacy evaluation of speech and natural language processing applications.

Diagnostic Evaluation

In speech and natural language processing application areas where coverage is important, for example in machine translation or language understanding systems with explicit grammars, a common development methodology employs a large *test suite* of exemplary input, whose goal is to enumerate all the elementary linguistic phenomena in the input domain, and their most likely and/or important combinations. A large, mature test suite will be structured into a number of dimensions of elementary phenomena and contexts, and may include invalid as well as valid inputs, tagged as such. Nerbonne, Netter, et al. (1993) describes a recent state-of-the-art example of this.

Test suites are particularly valuable to system developers and maintainers, allowing automated regression testing to ensure that system changes have the intended effect *and no others*, but raw profiles of system coverage *vis a vis* some test suite are unlikely to be of use as such in either adequacy or performance evaluation, both because such test suites may not reflect the distribution of linguistic phenomena in actual application domains, and because the *value* of good coverage at one point versus bad coverage at another is not in itself indicative of fitness to a user's purpose.

Performance Evaluation

There is a long tradition of quantitative performance evaluation in information retrieval, and many of its concepts have been usefully imported into the development of evaluation methodologies for speech and natural language processing. In particular, in considering any attempt at performance evaluation, we can usefully distinguish between three levels of specificity:

- **Criterion:** What it is we're interested in evaluating, in the abstract: Precision, Speed, Error rate
- **Measure:** Which specific property of system performance we report in an attempt to get at the chosen criterion: Ratio of hits to hits plus misses, seconds to process, percent incorrect.
- **Method:** How we determine the appropriate value for a given measure and a given system: Typically some form of concurrent or post-analytic measurement of system behavior over some benchmark task.

For example, in information retrieval itself, a classic criterion is precision, the extent to which the set of documents retrieved by a formal query satisfy the need which provoked the query. One measure for this is the percentage of documents retrieved which are in fact relevant. One method for computing this, which applies *only* if the extensions of some set of needs over some test collection are *known* in advance, is to simply average over some number of test queries the ratio achieved by the system under test.

For speech recognition, where the criterion is recognition accuracy, one measure is word error rate, and the method used in the current ARPA speech recognition evaluation involves comparing system transcription of the input speech to the *truth* (i.e., transcription by a human expert), using a mutually agreed upon dynamic programming algorithm to score agreement at the word level.

It should be clear from this that the distinction between criterion, measure and method is not hard and fast, and that in any given case the three are interdependent—see Sparck Jones (1994) for a more detailed discussion of these issues.

13.1.2 The Successes and Limitations of Evaluation

As the previous discussion illustrates, evaluation plays an important role for system developers (to tell if their system is improving), for system integrators (to determine which approaches should be used where) and for consumers (to identify which system will best meet a specific set of needs). Beyond this, evaluation plays a critical role in guiding and focusing research.

Periodic performance evaluations have been used successfully in the U.S. to focus attention on specific hard problems: robust information extraction from text, large vocabulary continuous speech recognition, spoken language interfaces, large scale information retrieval, machine translation. These *common evaluations* have motivated researchers both to compete in building advanced systems, and to share information to solve these hard problems. This paradigm has contributed to increased visibility for these areas, rapid technical progress, and increased communication among researchers working on these common evaluations as a result of the *community of effort* which arises from working on a common task using common data.

A major side-effect of performance evaluation has been to increase support for infrastructure. Performance evaluation itself requires significant investment to create annotated corpora and test sets, to create well-documented test procedures and programs, to implement and debug these procedures, and to distribute these to the appropriate parties.

Of course, the focus on performance evaluation comes at a price: periodic evaluations divert effort from research on the underlying technologies, the evaluations may emphasize some aspects of development at the expense of other aspects (e.g., increased accuracy at the expense of real-time interaction), and performance evaluation across systems can be misleading, depending on level of effort in developing the systems under comparison, use of innovative *vs.* proven technologies, and so on.

The *common evaluations* referred to above have all relied on performance evaluation, in part because some of them have received funding through ARPA, which focused on *technology* rather than on *applications*. Increasing emphasis on adequacy evaluation may become appropriate if, as seems likely on both sides of the Atlantic, users and their needs come more to the forefront of funding priorities. There is a difficulty in Europe, however, in that the basic performance evaluation technologies for languages other than English are developed only to a limited extent, with considerable variation across languages.

Successes of Evaluation

As noted above, evaluation has contributed some major successes to the development of speech and natural language processing technology; among these we can count:

- Development of test corpora for speech, spoken language, written language, information retrieval, and machine translation, and corresponding performance evaluation methods for aspects of these technologies. In addition, there are other shared resources, described in chapter 12.
- Creation of at least four performance evaluation conferences or workshops that have attracted increasing numbers of researchers, industry and government participants: the Message Understanding Conferences (MUCs), the Text Retrieval Conferences (TRECs), the Machine Translation Evaluation Workshops, and the Spoken Language Technology Workshops.
- Rapid technical progress: Evaluation allows progress to be tracked over time; for example, the word error rate for speech recognition has decreased by a factor of two every two years, over the last six years. Also, availability of performance evaluation methods make it possible to explore new paradigms based on automated learning algorithms, as in the use of parse evaluation techniques to build parsers (Brill, 1992).

Limitations of Current Evaluation Methods

As noted above, current evaluation technology also has some significant shortcomings and gaps:

- There has been little focus on how the user interacts with a system. Specifically, there is no performance evaluation methodology for interactive systems, and the methodologies for adequacy evaluation (and for formative evaluation) are difficult to apply and not widely accepted.
- Many of the performance evaluation methods are application-specific—that is, they require that everyone build the same application in order to evaluate their system (or their system components). We have yet to develop good methods of evaluating *understanding* independent of *doing the right thing* in the context of a specific application.
- There is no evaluation methodology for assessing how portable systems are to new application domains. The current high cost of porting language-based systems to

new applications hinders transition of this technology to the commercial marketplace.

- Evaluation is labor-intensive and and competes in time and resources with other activities, specifically with the development of new technical approaches. It is critical to find the right balance between technology development and technology assessment.

- Excessive focus on performance evaluation may lead to risk-avoidance strategies, where getting a good score becomes more important than doing good research. Evaluation must be counter-balanced by rewarding risk-taking, if research is to retain its vitality.

- Insufficient attention has been paid to evaluation in multilingual settings. In machine translation, there has been work on evaluation of different language pairs (English-French *vs.* English-Spanish *vs.* English-Japanese). There has also been work in text extraction for multiple languages (Japanese, Spanish). However, there is still a disproportionate emphasis on English, which presents a serious impediment to the widespread applicability of performance evaluation in Europe.

13.1.3 Future Directions

It is clear from both the successes and the shortcomings that evaluation methodologies will continue to evolve and to improve. Evaluation has become so central to progress in the speech and natural language area that it should become a research area in its own right, so that we can correct the problems that have become increasingly evident, while continuing to reap the benefits that evaluation provides.

13.2 Task-Oriented Text Analysis Evaluation

Beth Sundheim

Naval Command, Control and Ocean Surveillance Center RDT&E Division (NCCOSC/NRaD), San Diego, California, USA

The type of text analysis evaluation to be discussed in this section uses complete, naturally-occurring texts as test data and examines text analysis technology from the outside; that is, it examines technology in the context of an application system and treats the system as a black box. This type of evaluation is in contrast with ones that probe the internal workings of a system, such as ones that use constructed test suites of sentences to determine the coverage of a system's grammar. Two types of task-oriented text processing system evaluations have been designed and carried out on a large scale over the last several years:

1. Text retrieval has been evaluated in the context of:
   - a document routing task, where the system is tuned to match a statement of a user's persistent information need against previously unseen documents;
   - an ad hoc retrieval task, where the system is expected to match a user's one-time query against a more or less static (previously seen) text database.
2. Text understanding has been evaluated in the context of an information extraction task, where the system is tailored to look for certain kinds of facts in texts and to represent the output of its analysis as a set of simulated database records that capture the facts and their interrelationships. More recently, evaluations have been designed that are less domain-dependent and more focused on particular aspects of text understanding.

The forums for reporting the results of these evaluations have been the series of Text REtrieval Conferences (TREC) (Harman, 1993; Harman, 1994) and Message Understanding Conferences (MUC), particularly the more recent ones (DARPA, 1991b; DARPA, 1992b; ARPA, 1993b). The TRECs and MUCs are currently sponsored by the U.S. Advanced Research Projects Agency (ARPA) and have enjoyed the participation of non-U.S. as well as U.S. organizations.

The methodology associated with evaluating system performance on information extraction tasks has developed only in recent years, primarily through the MUC evaluations, and is just starting to mature with respect to the selection and exact formulation of metrics and the definition of readily evaluable tasks. In contrast, text retrieval evaluation methodology is now quite mature, having enjoyed over thirty years of development especially in the U.K. and U.S., and has been further developed via the TRECs, which have made substantial contributions to the text retrieval corpus development methodology and to the definition of evaluation metrics. With a fairly stable task definition and set of metrics, the TRECs have been able to measure performance improvements from one evaluation to the next with more precision than has so far been possible with the MUCs.

There are many similarities between TREC and MUC, including the following:

- Inclusion of both ARPA-sponsored research systems and other systems in the evaluations. Participation has included sites from North America, Europe, Asia and Australia.
- Use of large, naturally-occurring text corpora.
- Objective of end-to-end (*black box*) performance assessment.
- Evaluation metrics that are notionally similar, though different in formulation, reflecting the differences in the nature of the tasks.

The most enduring metrics of performance that have been applied to text retrieval and information extraction are termed *recall* and *precision*. These may be viewed as judging effectiveness from the application user's perspective, since they measure the extent to which the system produced all the appropriate output (recall) and only the appropriate output (precision). In the case of text retrieval, a correct output is a relevant document; in information extraction, a correct output is a relevant fact.

```
Recall = #relevant-returned/#relevant
Precision = #relevant-returned/#returned
```

In the above formulas, *relevant* refers to relevant documents in retrieval and to relevant facts in extraction; *returned* refers to retrieved documents in text retrieval and to extracted facts in information extraction. As will be explained below, text retrieval and information extraction represent fundamentally different tasks; therefore, the implementation of the recall and precision formulas also differs. In particular, the formulation of the precision metric for information extraction includes a term in the denominator for the number of *spurious* facts extracted, as well as the number of correct and incorrect facts extracted.

Typically, text retrieval systems are capable of producing ranked results, with the documents that the system judges more likely relevant ranked at the top of the list.

Evaluation of the ranked output results in a recall-precision curve, with points plotted that represent precision at various recall percentages. Such a curve is likely to show very high precision at 10% recall, perhaps 50% precision at 50% recall (for a challenging retrieval task), and a long tail-out toward 100% recall.

A simple information extraction task design might involve a fixed number of data elements (attributes) and a fixed set of alternative values for each attribute. If the system was expected always to produce a fixed number of simulated database records (sets of attributes), and a fixed number of facts per attribute from a fixed set of possible facts, it would be performing a kind of classification task, which is similar to the document routing task. In the document routing task performed by text retrieval systems, the routing queries represent categories, and the task is to determine which, if any, category is matched by a given text. However, an information extraction task typically places no upper bound on the number of facts that can be extracted from a text—the number of facts could conceivably even exceed the number of words in the text. In addition, a given fact to be extracted is not necessarily drawn from a predetermined list of possibilities (categories) but may instead be a text string such as the name of a victim of a kidnapping event.

Thus, since texts offer differing amounts of relevant information to be extracted and the *right answers* often do not come from a closed set, it is probably impossible for an information extraction system to achieve 100% recall except on the most trivial tasks, and its false alarms are likely to include large amounts of spurious data (as well simply erroneous data) if it is programmed to behave aggressively, in an effort to enable it to miss as little relevant information as possible. Current information extraction systems are not typically based on statistical algorithms, although there are exceptions. Therefore, evaluation typically does not produce a recall-precision curve for a system, but rather a single measure of performance.

One of the major contributions of both the TREC and the MUC evaluations has been the use of test corpora that are large enough to yield statistically valid performance figures and to support corpus-based system development experiments. The TREC-1 collection contained 200 times the number of documents found in a prior standard test collection (Harman, 1993). The MUCs have gradually brought about a similar revolution in the area of information extraction, which started in 1987 with a combined training and test corpus numbering just a few hundred, very short texts, and now uses several thousand longer texts; the number of test articles has increased from tens to hundreds.

To judge the correctness of the retrieval and extraction system outputs, the outputs must be compared with *ground truth.* Ground truth is determined by humans. In text retrieval, where the system may be evaluated using corpora consisting of tens of thousands of documents, it would be almost literally impossible to judge the relevance of all documents with respect to all queries used in the evaluation. Instead, one effective method in a multisystem evaluation on a corpus of that size is to pool the highest-ranked documents returned by each system and to judge the relevance of just those documents. For TREC-3, the 200 highest-ranked documents were pooled. It has been shown that different systems produce significantly different sets of top-ranking documents, and the pooling method can be fairly certain to result in a reasonably complete list of relevant documents (perhaps over 80% complete, on average across queries).

Information extraction systems have been evaluated using relatively small corpora (perhaps 100-300 documents) and just one or two extraction tasks. Ground truth is created by manually generating the appropriate database records for each document in the test set. Ground truth is not perfect truth in either retrieval or extraction, due not only to human factors but also to incomplete evaluation task explanations provided by the evaluators and to the inherent vagueness and ambiguity of text.

Widely varying system architectures, processing techniques, and tools have been tried, tested, and refined in the context of the MUC and TREC evaluations, accelerating progress in the robust processing of naturally-occurring text. There have been exciting innovations in technologies, including hybrid statistical/symbolic techniques and refined pattern-matching techniques. The infrastructure provided by the conferences and evaluations—shared corpora, evaluation metrics, etc.—and the conferences encourage the interchange of ideas and software resources and help participants understand which techniques work.

The need to isolate system strengths and weaknesses is one of the motivations underlying recent TREC and MUC efforts. These efforts have resulted in a greater range of evaluation options for participants. For example, the range of MUC evaluations has broadened from a single, complex, domain-dependent information extraction task to include also a simple, domain-independent task, and other tasks have been developed to test component-level technologies, such as identification of coreference relations and recognition of special lexical patterns such as person and company names. Various corporate and government organizations in Europe and the U.S. have sponsored similar component-technology, multisite evaluation efforts. These have focused especially on grammars and morphological processors as, for example, did the 1993 Morpholympics evaluation, coordinated by the Gesellschaft für Linguistische Datenverarbeitung (Hausser, 1994).

13.3 Evaluation of Machine Translation and Translation Tools

John Hutchins

University of East Anglia, Norfolk, UK

While there is general agreement about the basic features of machine translation (MT) evaluation (as reflected in general introductory texts Lehrberger & Bourbeau, 1988; Hutchins & Somers, 1992; Arnold et al., 1994), there are no universally accepted and reliable methods and measures, and evaluation methodology has been the subject of much discussion in recent years (e.g., Arnold et al., 1993; Falkedal, 1994; AMTA, 1992).

As in other areas of NLP, three types of evaluation are recognised: adequacy evaluation to determine the fitness of MT systems within a specified operational context; diagnostic evaluation to identify limitations, errors and deficiencies, which may be corrected or improved (by the research team or by the developers); and performance evaluation to assess stages of system development or different technical implementations. Adequacy evaluation is typically performed by potential users and/or purchasers of systems (individuals, companies, or agencies); diagnostic evaluation is the concern mainly of researchers and developers; and performance evaluation may be undertaken by either researchers/developers or by potential users. In the case of production systems there are also assessments of marketability undertaken by or for MT system vendors.

MT evaluations typically include features not present in evaluations of other NLP systems: the quality of the *raw* (unedited) translations, e.g., intelligibility, accuracy, fidelity, appropriateness of style/register; the usability of facilities for creating and updating dictionaries, for post-editing texts, for controlling input language, for customisation of documents, etc.; the extendibility to new language pairs and/or new subject domains; and cost-benefit comparisons with human translation performance. Adequacy evaluations by potential purchasers usually include the testing of systems with sets of *typical* documents. But these are necessarily restricted to specific domains, and for diagnostic and performance evaluation there is a need for more generally applicable and objective *test suites*; these are now under development (King & Falkedal, 1990; Balkan et al., 1994).

Initially, MT evaluation was seen primarily in terms of comparisons of unedited MT output quality and human translations, e.g., the ALPAC evaluations (Council, 1966) and those of the original Logos system (Sinaiko & Klare, 1972; Sinaiko & Klare, 1973). Later, systems were assessed for quality of output and usefulness in operational contexts, e.g., the influential evaluations of Systran by the European Commission (Van Slype, 1982). Subsequently, many potential purchasers have conducted their own comparative evaluations of systems, often unpublished, and often without the benefit of previous evaluations. Valuable contributions to MT evaluation methodology have been made by Rinsche (1993) in her study for the European Commission, and by the JEIDA committee (Nomura & Isahara, 1992), which proposed evaluation tools for both system developers and potential users—described in more detail in section 13.5. The evaluation exercise by ARPA (White et al., 1994) compared the unedited output of the three ARPA-supported experimental systems (Pangloss, Candide, Lingstat) with the output from 13 production systems from Globalink, PC-Translator, Microtac, Pivot, PAHO, Metal, Socatra XLT, Systran, and Winger. The initial intention to measure the *productivity* of systems for potential users was abandoned because it introduced too many variables. Evaluation, therefore, has concentrated on the performance of the *core MT engines* of systems, in comparison with human translations, using measures of adequacy (how well a text *fragment* conveys the information of the source), fluency (whether the output reads like good English, irrespective of accuracy), and comprehension or informativeness (using SAT-like multiple choice tests covering the whole text).

Future Directions

With the rapid growth in sales of MT software and the increasing availability of MT services over networks there is an urgent need for MT researchers, developers and vendors to agree and implement objective, reliable and publicly acceptable benchmarks, standards and evaluation metrics.

13.4 Evaluation of Broad-Coverage Natural-Language Parsers

Ezra Black

Interpreting Telecommunications Laboratories, ATR, Kyoto, Japan

13.4.1 State of the Art

A *parser* for some natural language (English, Portugese, etc.) is a program that *diagrams* sentences of that language—that supplies for a given sentence a correct grammatical analysis, demarcating its parts (called *constituents*), labeling each, identifying the part of speech of every word used in the sentence, and usually offering additional information, such as the *semantic class* (e.g., Person, Physical Object) of each word and the *functional class* (e.g., Subject, Direct Object) of each constituent of the sentence. A *broad-coverage parser* diagrams *any* sentence of some natural language, or at least agrees to attempt to do so.

Currently the field of broad-coverage natural-language parsing is in transition. Rigorous, objective and verifiable evaluation procedures have not yet become established practice, although a beginning has been made. Until recently, objective evaluation essentially was not practiced at all, so that even the author of a parsing system had no real idea how accurate, and hence how useful, the system was. In 1991 the Parseval system for *syntactically* evaluating broad-coverage English-language parsers was introduced (Black, Abney, et al., 1991; Harrison, Abney, et al., 1991), and the next year seven creators of such parsers applied Parseval to their systems, all using the same test data (Black, Garside, et al., 1993).

However, the Parseval evaluation routine is an extremely coarse-grained tool. For one thing, most of the information provided by a parse is not taken into account. But more importantly, the level of agreement on the particulars of linguistic description is fairly superficial among the creators of Parseval, and a fortiori among parsing-system authors who could or would not be included in the Parseval planning sessions. Consequently, parsers are evaluated by Parseval at a high remove from the actual parses being judged, and in terms rather foreign to their own vocabulary of linguistic description.

Currently there are plans to extend Parseval into the *semantic* realm, via Semeval, an approach to evaluation modeled on Parseval (see Moore, 1994). But there is *more, not less* disagreement among professionals regarding the proper set of semantic categories for text, the various word senses of any given word, and related semantic issues, than there is about constituent boundaries. So Semeval can be expected to turn out even rougher-grained than Parseval.

13.4.2 Improving the State of the Art

The methodology of objective, rigorous, and verifiable measurement of performance of individual parsing systems is known, albeit by only a minority of practitioners. Key features of this methodology are the use of:

1. *separate* training and test sets;
2. test data from *new* documents only;
3. *large* test sets;
4. responsible *public access* to the test process;
5. *objective* criteria of evaluation;
6. the statement, in advance, of *all* acceptable analyses for a test item;
7. test runs on a *variety* of test materials to match the sort of claims being made for the system; and
8. at least a *twice-yearly* run of a full range of public tests.

A slow transition is now taking place within the field towards the recognition of the value, and even the necessity, of rigor of the above sort within evaluation. This kind of testing is necessary anyway for effective parsing-system development, as opposed to the onerous activities associated with testing via *compromise-based* tools such as Parseval, Semeval, or others. It may never be possible to compare *all* broad-coverage parsers of a given language in terms of a *common coin* of linguistic analysis. Instead, practitioners will probably want to opt for highly accurate and rigorous performance statistics on their own systems alone, rather than extremely coarse-grained scores obtained from comparing their systems with others on the basis of laborious and even dubious technical compromise.

Another progressive development has been the appearance since 1992 of parsing systems which parse previously-unseen text without referring to a set of grammar rules, by processing, statistically or logistically, a *treebank* or set of sentences parsed correctly by hand by competent humans (Black, 1993). These systems are in theory directly comparable, and can employ more rigorous correctness criteria—e.g., exact match of the treebank parse—than can Parseval.

13.4.3 Future Directions

The remainder of the 1990s will probably see two major trends in this area. First should be a move toward the sort of rigor discussed above, when individual systems are evaluated either just to let the system developer himself or herself know the rate at which and the manner in which the system is improving over time, or else for the purpose of cross-system comparisons on a given document, where this is possible (see above). Second should be a move away from evaluating parsing systems in linguistic terms at all, i.e., away from judging the parses output by a system simply on their merits as parses. This move would be *toward* evaluating a parser on the basis of the *value added* to a variety of *client systems*. These would be bona fide, fully-developed AI systems of one sort or another, with a need for a parsing component. This as opposed to tasks conceived artificially, simply for the purposes of providing a *task* to support evaluation. Examples might be pre-existing systems for speech synthesis, speech recognition, handwriting recognition, optical character recognition, and machine translation. In this case the evaluation of a broad-coverage parsing system would come to be based on its performance over a *gamut* of such applications.

13.5 Human Factors and User Acceptability

Margaret King

University of Geneva, Switzerland

It is quite astonishing how little attention is paid to users in the published literature on evaluation. To some extent, this can be explained by looking at who does evaluation and is prepared to talk about it. Essentially, we find three classes:

- Researchers or manufacturers concerned with system development: The researchers do not have the resources to carry out any systematic enquiry into what a group of users might actually want. The developers mainly come into contact with users through their customer support services. In both cases, when a user is taken into account, it is an abstract, ideal user, whose needs correspond to those the researcher or system developer thinks he would have.
- Funding agencies, especially, in this context, ARPA: Since what they are primarily interested in is the development of a core technology, evaluation is seen as an assessment of a system's ability to perform a pre-determined task taken to reflect the barriers the core technology should be attacking. In this perspective, thinking of an ultimate user is premature and irrelevant.
- Potential purchasers of commercially available systems: Here, of course, the user is directly present, but concerned only with his own needs.

13.5.1 State of the Art

One exception to the above comes from the area of machine translation. The Japan Electronic Industry Development Association's Machine Translation System Research Committee has a sub-committee, the Machine Translation Market and Technology Study Committee, which has recently published a report on evaluation criteria for machine translation systems. (A summary account can be found in: Nomura & Isahara, 1992.)

The committee concentrated on three aspects:

- **User Evaluation of Economic Factors:** The aim is to support making decisions about what kind of system is suitable in those cases where introducing a machine translation system in the near future is being considered. Economic factors only are taken into consideration.

- **Technical Evaluation by Users:** The aim is to compare the users' needs with what is offered by a particular system, rather than to offer any abstract evaluation of the system per se.
- **Technical Evaluation by Developers:** The aim here is to support in-house evaluation of the technical level the system has achieved and of whether the system suits the purpose for which it was developed.

In what follows we shall concentrate on the first two aspects:

User evaluation of economic factors is essentially accomplished by analysing the replies to two questionnaires, the first concerning the user's present situation, the second his perceived needs. The answers are evaluated in the light of a set of parameters relating the answers to what advantages a machine translation system could offer. The results of are presented graphically in the form of a *radar chart*, which provides a profile of the user.

In parallel, a similar exercise is carried out to produce profiles of typical users of types of machine translation systems. Seven types of systems are distinguished in all, which cover in fact the whole range of translators' aids. The committee members define a typical user for each type of system, and a profile for that user is constructed on the basis of the answers he would be expected to give to the questionnaire. This profile then becomes the profile of the system-type. Types of system can then be paired with types of users by comparing the radar chart profiles for user and for system and finding the closest match.

The validity of the procedure is confirmed by taking, for each system type, a further group of four (assumed) users, filling out the questionnaires on their behalf, and checking that the closest match is what is expected to be.

Two points are worth making about this procedure. The first is that what is being considered is not really systems but what Galliers and Sparck Jones (1993) call *setups*, that is, a system embedded in a context of use. This is important: from a real user's point of view, there is usually very little point in evaluating a system in isolation. The ISO 9000 series on quality assessment of software makes the same point, although from a rather different viewpoint:

> "The importance of each quality characteristic varies depending on the class of software. For example, reliability is most important for a mission critical system software, efficiency is most important for a time critical real time system software, and usability is most important for an interactive end user software."—ISO (1991)

The second point shades rather to the negative; the users considered in constructing the radar charts of the system type are not real users. It is important to be aware of the dangers involved in deciding on behalf of some third party what it is he really wants or needs.

This potential weakness is partially at least counterbalanced by the second type of evaluation, called in the committee's reports "technical evaluation by users." Here, an attempt is made to determine the user's real needs and to compare them with what can be offered by specific products in order to evaluate how satisfied the client is likely to be with what is offered.

Attempts to take user needs into consideration were also made within the Esprit Translators' Workbench projects (ESPRIT project 2315, TWB I and 6005, TWB II). Catalogues were developed for describing user requirements, term banks, translation memories, machine translation, machine assisted terminology work and for checkers. The catalogues were intended to serve a double purpose, first as a way of setting up requirements specifications, and secondly as a way of evaluating to what extent a particular tool corresponds to a given user's needs. In general terms, each catalogue comprises facts relevant to the software and related to a certain quality characteristic, such as task adequacy, error tolerance, execution efficiency, ease of use, ease of learning, etc. Users can tick items which are relevant to them, give items an individual priority and rate each priority by specifying its relative importance compared to other items of the same type (Höge, Hohmann, et al., 1992; Höge, Hohmann, et al., 1993).

13.5.2 Current Work

In this section, we look at the efforts of the EAGLES Evaluation Group to build on these and other efforts in order to define an evaluation methodology where the users' views and needs are systematically taken into account.

The overall aim of the Evaluation Group is to define a common general framework within which specific evaluations can be designed. In this work it has also been influenced by the discussions reported in Thompson (1992), by the work of Galliers and Sparck Jones (1993) and by the work on evaluation within the ARPA/DARPA community.

The group distinguishes three types of evaluation: progress evaluation, where the aim is to assess the progress of a system towards some other ideal state of the same system, diagnostic evaluation, where the aim is to find out where things go wrong and why, and adequacy evaluation, where the aim is to assess the adequacy of the system to fulfill a specified set of needs.

User-centered evaluation is clearly adequacy evaluation. The first problem becomes evident at this point. Adequacy evaluation involves finding out whether a product satisfies the user's needs. But users are very numerous, and have widely differing needs. It would be out of the question to work in terms of individuals. However, on the basis of surveying what a sufficiently large number of individual users say, it should be possible to identify classes of users and to construct profiles of each one of these classes. These profiles can then be used as the basis for determining what attributes of particular classes of products are of interest to particular classes of users. Then, for each such attribute, a procedure can be specified for discovering its value in the case of any particular product.

The appropriate analogy is with the kind of reports published by consumer associations, where different products of the same general class are compared along a number of different dimensions. Consumer reports typically are concerned with products based on a relatively stable technology. Transferring the paradigm to the more sophisticated products of the language industry can require a great deal of work, and sometimes a considerable degree of ingenuity. In the interest of producing concrete results in the short term, while at the same time checking the validity of the general framework, the EAGLES group, together with an associated LRE project, TEMAA, is concentrating on designing evaluation packages for market or near market products in two areas, authoring aids and translation aids. These areas are of particular interest partly because the market is large, and therefore the results are likely to be of interest to a large number of potential users, partly because at least some of the products in these areas are based on a fairly stable technology.

If it proves possible to produce evaluation packages for a range of language industry products, they can be expected to constitute a de facto standard for such products. Working on how this can be done for the more modest products of the language industry lays the foundation for extending the enterprise to more sophisticated products.

13.6 Speech Input: Assessment and Evaluation

David S. Pallett[a] & Adrian Fourcin[b]

[a] National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[b] University College of London, London, UK

Assessment and evaluation[1] are concerned with the global quantification and detailed measurement of system performance. Disciplined procedures of this type are at the heart of progress in any field of engineering. They not only make it possible to monitor change over time in a given system and meaningfully compare one approach with another; they also usefully extend basic knowledge.

Within the past several years, there has been widespread and growing international interest in a number of issues involved in speech input system performance assessment. In Europe, the SAM Projects (ESPRIT Projects 2589 and 6819) addressed "Multi-Lingual Speech Input/Output Assessment, Methodology and Standardization" (Fourcin et al., 1992). In the United States, the ARPA Spoken Language Program has made extensive use of periodic *benchmark tests* to gauge progress and to serve as a focal point for discussions at a number of ARPA-sponsored workshops (DARPA, 1989; DARPA, 1990; DARPA, 1991a; DARPA, 1992a; ARPA, 1993a; ARPA, 1994).

There have also been a number of international workshops, such as those held in conjunction with the Eurospeech Conferences and the International Conferences on Spoken Language Processing (Jones & Mariani, 1992). The present contribution focuses on three sub-areas: speech recognition; speech understanding; and speaker recognition.

13.6.1 Speech Recognition (Input) Assessment

To a first approximation, the task of *speech recognition* may be regarded as being to produce an hypothesized orthographic transcription from a spoken language input. The most commonly cited output is in the form of *words* in ASCII characters, although other units (e.g., syllables or phonemes) are sometimes found.

Assessment methods developed for speech recognition involve a complementary combination of system based approaches with performance based techniques. System based approaches either deal with the recognition system as a whole (*black box methods*) or provide access to individual modules within the complete recognizer (*glass box methods*). For each of these approaches quantitative appraisals of performance may range from the use of applications related (non-diagnostic) training and test data to

[1] See the appendix at the end of this section for a definition of *assessment* and *evaluation* within SAM.

highly diagnostic techniques, specifically oriented toward detailed evaluations involving the use of test data going from, for example, phonetically controlled speech to language independent data derived from artificial speech generation. These extremes of performance measurement fit into a continuum into which methods of global, benchmarking, assessment and detailed evaluation may be categorized into the following groups:

(a) application-oriented techniques based on the use of general databases, collected under what might be regarded as representative conditions

(b) the use of specific calibrated databases, which are designed to represent a broad spectrum of operational and environmental conditions which affect recognizer performance

(c) the use of reference methods in order to achieve cross-site standardization, based on the use of a reference recognizer, or referring to human recognition

(d) diagnostic methods, based on the use of specific vocabularies or specially designed sequences

(e) techniques using artificial test signals to achieve precision of control of the experimental design and/or language independence

(f) benchmarking which is based on predictive methods using system parameters and/or speech knowledge

The most frequently used methods (e.g., those used within the ARPA programme) belong to group (a). Much of the data used for speech recognizer performance assessment consists of *read* speech, not spontaneous, goal-directed speech. Some of the data used for large-scale performance assessment efforts is openly available (see section 12.6).

Automatic scoring methods are used in most cases, with reliance on dynamic programming methods to align reference and system hypothesis output strings. Results are typically reported in terms of the word or sentence error percentages, where errors are categorized as substitutions, insertions, or deletions.

The statistical validity of assessment tests for recognizers has been studied (Chollet, Capman, et al., 1991), and a number of well-known statistical measures are in use, using both parametric and non-parametric techniques.

13.6.2 Speech (Spoken Language) Understanding Assessment

In *speech understanding* some semantic analysis or interpretation of the speech recognizer's output is implicitly or explicitly required—for example where the process of automatic speech recognition is intended as input to a command/control application.

Performance assessment for speech, or more generally *spoken language*, understanding systems is substantially more complex and problematic than for speech recognition systems. Procedures for performance assessment of natural language processing systems, in general, are not yet well established, but many relevant issues have been identified and addressed in increasing detail at workshops in Pennsylvania in 1988, Berkeley in 1991, Edinburgh and Trento in 1992, as well as at the ARPA Human Language Technology Workshops.

For spoken language understanding systems, the use of reference speech databases as system input is not so clearly appropriate, because issues involving human behavior and human-computer interactivity become complicating factors. "It is particularly difficult to engage in speech evaluation where the entire system design assumes a high degree of interaction between user and system, and makes explicit allowance for [dialogue] clarification and recovery, as in the VODIS telephone train inquiry case" (Galliers & Sparck Jones, 1993).

Nonetheless, this procedure has, for example, extensively been implemented within the ARPA Spoken Language Program in the U.S., in the Air Travel Information Service (ATIS) domain, a spoken natural language (air travel information) database query task. "The evaluation methodology is *black box* and implemented using an automatic evaluation system. It is performance related; only the content of an answer retrieved from the database is evaluated" (Galliers & Sparck Jones, 1993).

A variety of procedures have been suggested for accommodating interactive systems with dialogue management and/or clarification. So-called *end-to-end* assessment methods—in which measures of system-user efficiency in task completion and/or subjective measures of satisfaction are derived—are frequently complicated by large subject-to-subject or task-to-task variabilities, and their attendant statistical considerations. It is clear that these complications will be relevant to the assessment and benchmarking of commercial technology for real applications, as well as to their detailed evaluation and future development.

13.6.3 Speaker Recognition Assessment

Speaker recognition technology is conventionally discussed in terms of two different areas: speaker identification and speaker verification (see section 1.7). Speaker identification can often be thought of as a *closed set* problem, where the system's task is to identify an unidentified voice as coming from one of a set of N reference speakers. In practical applications, *open set* speaker identification permits a rejection response corresponding to the possibility that the unidentified voice does not belong to any of the reference speakers. The task of a speaker verification system is to decide whether the unlabeled voice belongs to a specific *genuine* speaker who has previously claimed his identity to the system, or an *imposter*.

A state-of-the-art in the evaluation of speaker identification and verification systems can be found in the Proceedings of the Automatic Speaker Recognition, Identification and Verification ETRW Workshop (Chollet, Bimbot, et al., 1994), et al., 1994), as a summary of the initial efforts of ESPRIT Project 6819, Speech Technology Assessment Methodology in Multilingual Applications (SAM-A) (Bimbot et al., 1994). et al.,).

13.6.4 Future Directions

In the shorter term, provision should be made for more accurate speech recognition scoring procedures making use of time-marked reference transcriptions and system outputs. Such procedures may prove essential when conducting multi-lingual performance assessment, to facilitate cross comparison and, for example, because of increased ambiguity concerning word boundaries for some languages. The adequate provision of these facilities will involve quite new approaches to the large scale accurate labeling of speech databases.

The increasingly wide area of applications of speech recognition technology introduces new needs and new problems. The need to support fluent dialogue interaction with a range of speakers, accents, dialects and conditions of health increases the complexity of assessment and evaluation for developer and user alike. For truly spontaneous speech input collected in operational environments, the presence of disfluencies (e.g., pause-fillers, word fragments, false starts and restarts) and noise artifacts provide additional complicating factors.

The associated need for systems to be able to be trained so as to work with a range of language inputs similarly imposes a much greater burden on the organization and collection of appropriate spoken language corpora. This in turn should lead to the gradual use of more analytic and language independent techniques (glass box techniques) and an increasingly close association between work in speech input with speech output/synthesis and natural language processing.

Appendix: Assessment and Evaluation Defined

The increasing complexity of processing associated with the development and application of spoken language processing systems is necessarily tied in with an increasing need for precision, both in the methods employed for the appraisal of performance and in our use of the description of these methods.

Assessment is the process of system appraisal which leads to global, overall, quantification of performance. Assessment is related conceptually to black box methods in which the detailed mechanisms of processing are not considered. (The word itself has its origin in the latin *assidere*—to sit by—and relates to the levying of tax on the gross production of an enterprise.)

Evaluation involves the analytic description of system performance in terms of defined factors, it is concerned with detailed measurement. Evaluation is conceptually related to the *glass box* approach, in which the objective is, for example, to gain a greater understanding of system performance from the use of precision diagnostic techniques based on special purpose phonetic databases. (The word itself has its origins in the French word *evaluer*—to calculate from a mathematical expression or to express in terms of something already known.)

13.7 Speech Synthesis Evaluation

Louis C. W. Pols

University of Amsterdam, The Netherlands

The possibility to generate any existing text, any to-be-worked-out concept, or any piece of database information as intelligible and natural sounding (synthetic) speech is an important component in many speech technology applications (Sorin, 1994). System developers, product buyers, and end users are all interested in having appropriate scores to specify system performance in absolute (e.g., percentage correct phoneme or word intelligibility scores) and in relative terms (e.g., this module sounds more natural for that specific application in that language than another module) (Jekosch, 1993).

Since synthetic speech is generally derived from text input (see also chapter 5), not just a properly functioning acoustic generator is required, but also proper text interpretation and preprocessing, grapheme-to-phoneme conversion, phrasing and stress assignment, as well as prosody, and speaker and style characteristics have to be adequate. On all these, and several other, levels one might like to be able to specify the performance, unless one really only wants to know whether a specific task can properly be performed in a given amount of time. This opposes the approach of modular diagnostic evaluation to the one in which global overall performance is the main aim.

13.7.1 Modular Diagnostic Evaluation

At this diagnostic level a suite of tests is already available, although there is little standardization so far, nor are there proper benchmarks. Also comparability of test design and interpretability of results over languages, is a major point of concern (Logan, Greene, et al., 1989; Pols, 1991). The type of tests we have in mind here are methods to evaluate system performance at the level of text pre-processing, grapheme-to-phoneme conversion, phrasing, accentuation (focus), phoneme intelligibility, word and (proper) name intelligibility (Spiegel, 1993), performance with ambiguous sentences, comprehension tests, and psycho-linguistic tests such as lexical decision and word recall. There is a great lack of proper tests concerning prosody, and speaker, style and emotion characteristics, but this is partly so because rule-synthesizers themselves are not yet very advanced concerning these aspects either (Pols, 1994b). However, concatenative synthesis with units taken from large databases plus imitation of prosodic characteristics, is one way to overcome this problem of insufficient knowledge concerning detailed rules. The result is high-quality synthesis for specific applications with one voice and one style only.

13.7.2 Global Overall Performance

In this global category fall the overall quality judgments, such as the mean opinion score (MOS), as commonly used in telecommunication applications. Such tests have little diagnostic value, but can clearly indicate whether the speech quality is acceptable for a specific application by the general public. One can think of telecommunication applications such as a spoken weather forecast, or access to e-mail via a spoken output. Also prototypes of reading machines for the visually-impaired, allowing them to listen to a spoken newspaper, are evaluated this way. In field tests not just the speech quality, but also the functionality of the application should be evaluated.

13.7.3 Towards International Standards

Although presently there is little standardization and proper multilingual benchmarks for speech synthesis are lacking, various organizations are working on it. Via the Spoken Language Working Group in Eagles, a state-of-the-art report with recommendations on the assessment of speech output systems has been compiled (Eagles, 1995), largely based on earlier work within the Esprit-SAM project (Pols & SAM-partners, 1992). The Speech Output Group within the world-wide organization COCOSDA has taken various initiatives with respect to synthesis assessment and the use of databases (Pols & Jekosch, 1994). One recent intriguing proposal is to arrange real-time access to any operational text-to-speech system via World Wide Web. The ITU-TS recently produced a recommendation about the subjective performance assessment of synthetic speech over the telephone (ITU, 1993; Klaus, Klix, et al., 1993).

13.7.4 Future Directions

In the future, we will probably see more and more integrated text and speech technology in an interactive dialogue system where text-to-speech output is just one of several output options (Pols, 1994a). The inherent quality of the speech synthesizer should then also be compared against other output devices such as canned natural (manipulated) speech, coded speech, and visual and tactile displays. Also the integration of these various elements then becomes more important, and their performance should be evaluated accordingly.

13.8 Usability and Interface Design

Sharon Oviatt

Oregon Graduate Institute of Science & Technology, Portland, Oregon, USA

To date, the development of spoken language systems primarily has been a technology-driven phenomenon. As speech recognition has improved, progress traditionally has been documented in the reduction of word error rates (Pallett, Fiscus, et al., 1994). However, reporting word error rate fails to express the frustration typically experienced by users who cannot complete a task with current speech technology (Rhyne & Wolf, 1993). Although the successful design of interfaces is essential to supporting usable spoken language systems, research on human-computer spoken interaction currently represents a gap in our scientific knowledge. Moreover, this gap is widely recognized as having generated a bottleneck in our ability to deploy robust speech technology in actual field settings.

Among other challenges, interfaces will be needed that can guide users' spontaneous speech to coincide with system capabilities, since spontaneous speech is known to be particularly variable along a number of linguistic dimensions (Cole, Hirschman, et al., 1995). Interface techniques for successfully constraining spoken input have been studied most extensively by the telecommunications industry as it strives to automate operator services (Karis & Dobroth, 1991; Spitz, 1991). Such work has emphasized the need for realistic and *situated* user testing, often in field settings, and has shown that dramatic variation can occur in the successful elicitation of target speech depending on the type of system prompt.

Other research has demonstrated that the principle of linguistic convergence, or the tendency of people's speech patterns to gravitate toward those of their interactive partner, can be employed to guide wordiness, lexical choice, and grammatical structure during human-computer spoken interactions, and without imposing any explicit constraints on user behavior (Zoltan-Ford, 1991). In addition, research has shown that difficult sources of variability in human speech (e.g., disfluencies, syntactic ambiguity) can be reduced by a factor of 2-to-8 fold through alteration of interface parameters (Oviatt, 1995; Oviatt, Cohen, et al., 1994). Such work demonstrates the potential impact that interface design can have on managing spoken input, although interface techniques have been underexploited for this purpose. In all of these areas, research typically has involved *proactive* performance assessment using simulation techniques, which is the preferred method of conducting evaluations of systems in the planning stages.

Future Directions

Many basic issues need to be addressed before technology can leverage fully from the natural advantages of speech—including the speed, ease, spontaneity, and expressive power that people experience when using it during human-human communication. For example, research is needed to evaluate different types of natural spoken dialogue, spontaneous speech characteristics and their management, and dimensions of human-computer interactivity that influence spoken communication. With respect to the latter, research is especially needed on optimal delivery of system confirmation feedback, error patterns and their resolution, flexible regulation of conversational control, and management of users' inflated expectations of the *interactional* coverage of spoken language systems. In addition, the functional role that ultimately is most suitable for speech technology needs to be evaluated further. Finally, assessment is needed of the potential usability advantages of multimodal systems incorporating speech over unimodal speech systems, with respect to breadth of utility, ease of error handling, learnability, flexibility, and overall robustness (Cohen & Oviatt, 1994; Cole, Hirschman, et al., 1995). To support all of these research agendas, tools will be needed for building and adapting high quality, semiautomatic simulations. Such an infrastructure can be used to evaluate the critical performance tradeoffs that designers will encounter as they strive to design more usable spoken language systems.

13.9 Speech Communication Quality

Herman J. M. Steeneken

TNO Human Factors Research Institute, Soesterberg, The Netherlands

Speech is considered to be the major means of communication between people. In many situations, however, the speech signal we are listening to is degraded, and only a limited transfer of information is obtained. The purpose of assessment is to quantify these limitations and to identify the limitations responsible for the loss in intelligibility. For assessment of speech communication systems mainly three major evaluation methods are used:

1. Subjective intelligibility based on scores for correct recall of sentences, words or phonemes;
2. quality ratings based on a subjective impression; and
3. objective measures based on physical properties of the speech transmission system.

A comprehensive overview is given by Steeneken (1992).

13.9.1 Subjective Intelligibility Tests

These are based on various types of speech material evaluated in speaker-listener communication. All these tests have their specific advantages and limitations, mostly related to the speech elements tested. Speech elements frequently used for testing are phonemes, words (digits, alphabet, meaningful words, or nonsense CVC-words (Consonant-Vowel-Consonant), sentences, and sometimes a free conversation. The percentage correctly recalled items of the set presented gives the score. The recall procedure can be based on a given limited set of responses or on an open response design in which all possible alternatives are allowed as a response. A limited response set is used with the so-called rhyme tests. These type of tests are easy to administer and do not require extensive training by the listeners in order to arrive at stable scores. Rhyme tests may, depending on the design, disregard specific phoneme confusions (House, Williams, et al., 1965). Open response tests, especially those which make use of nonsense words, require an extensive training of the listeners. However, additionally to the word and phoneme scores, possible confusions between phonemes are obtained. This allows for diagnostic analysis. Redundant speech material (sentences, rhyme tests) suffers from ceiling effects (100% score at poor-to-fair conditions) while tests based on nonsense words may discriminate between good and excellent conditions.

13.9.2 Quality Rating or Mean Opinion Scoring (MOS)

As noted in sections 10.2.2 and 10.2.3, MOS is a more global method used to evaluate the user's acceptance of a transmission channel or speech output system. It reflects the total auditory impression of speech by a listener. For quality ratings, normal test sentences or a free conversation are used to obtain the listener's impression. The listener is asked to rate his impression on subjective scales such as: intelligibility, quality, acceptability, naturalness, etc. The MOS gives a wide variation among listener scores and does not give an absolute measure since the scales used by the listeners are not calibrated.

13.9.3 Objective Measures

Objective measures based on physical aspects quantify the effect on the speech signal and the related loss of intelligibility due to deteriorations as: a limited frequency transfer, masking noises with various spectra, reverberation and echoes, and a nonlinear transfer resulting from peak clipping, quantization, or interruptions. Frequently used methods are the Articulation Index (AI) (Kryter, 1962) and the Speech Transmission Index (STI), (Steeneken & Houtgast, 1980). The STI makes use of artificial test signals which are passed through the system under test and analyzed at the output-side. Such a measurement can be performed typically in 15 seconds (Steeneken, Verhave, et al., 1993), while subjective measurements require at least one hour.

In Figure 13.1 the relation between some intelligibility measures and the STI is given. These results are based on cumulated results obtained over the years. A subjective qualification, based on an international comparison (Houtgast & Steeneken, 1984), is also given. The graph also demonstrates the ceiling effect of intelligibility tests making use of redundant speech material .

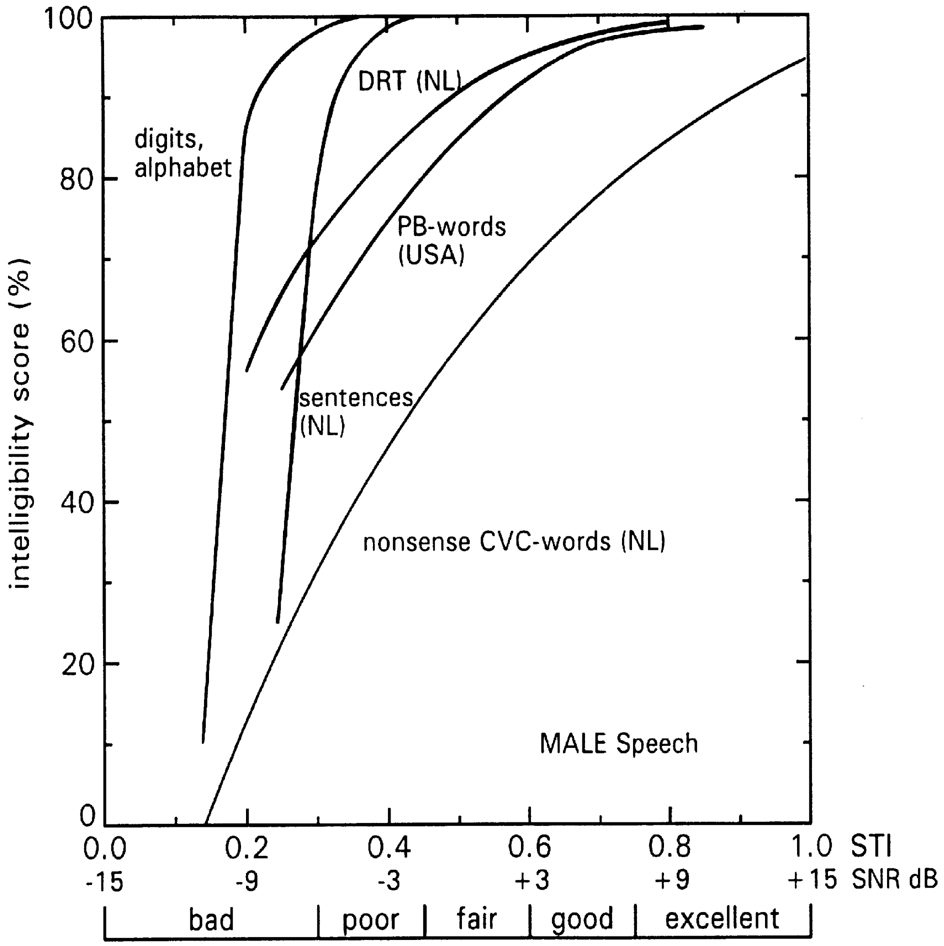

Figure 13.1: Relation between and qualification of some subjective intelligibility measures and the objective STI.

13.10 Character Recognition

Junichi Kanai

University of Nevada, Las Vegas, Nevada, USA

The variables that affect the performance of an optical character recognition (OCR) system include variations in the clarity of printed documents, as well as their layout style. These factors contribute to the number of needed performance metrics, to the need for large quantities of test data, and the necessity of automating the evaluation task.

Traditionally, the performance of OCR algorithms and systems is based on the recognition of isolated characters. When a system classifies an individual character, its output is typically a character label or a reject marker that corresponds to an unrecognized character. By comparing output labels with the correct labels, the number of correct recognition, substitution errors (misrecognized characters), and rejects (unrecognized characters) are determined. The standard display of the results of classifying individual characters is the confusion matrix, such as Figure 13.2.

```
                      Recognized as
                  a  b  c  d  | Reject | Error
                -------------------------------
               a  9        1  |        |   1
True ID        b     8        |   2    |   0
               c  2     6  1  |   1    |   3
               d  1        9  |        |   1
                -------------------------------
                  3  0  0  2  |   3    |   5
```

Figure 13.2: Confusion matrix

The character accuracy is:

Recognized-Characters / Input-Characters

The cost of correcting residual errors in output is:

$W_1 \times$ Substitution-Errors $+ W_2 \times$ Rejects where W_1 and W_2 are costs associated with correcting a substitution error and a reject, respectively.

Many OCR systems use morphological (n-gram) and lexical techniques to correct recognition errors. To evaluate the performance of such systems, word, sentence, or paragraph images are needed. Since linguistic characteristics, such as n-gram statistics and word frequency, depend on document class (or domain), standard lexicons or corpora for training and testing extracted from a variety of document classes are needed. As OCR systems employ other natural language processing techniques to improve accuracy, appropriate training and test databases must be developed.

OCR and document analysis systems recognize not only text but also other features of documents, such as extraction of articles from a page and recognition of the logical structure of an article. New metrics and appropriate resources, such as document-based test data must be made available.

Since the notion of *accuracy* depends upon the specific application involved, application-specific metrics are also important. Such metrics can also help end users to determine the feasibility of OCR in their tasks. Consider text retrieval applications. Users of text retrieval systems are interested in words and their correct reading order and almost never in individual characters. Thus, *word accuracy* is a more appropriate metric. Moreover, for these applications, discriminating between *stopwords* and *non-stopwords* is important. Stopwords are common words, such as *the*, *but*, and *which*, that are normally not indexed because they have essentially no retrieval value. Therefore, correct recognition of words that are not stopwords is an even more important metric for these applications (Rice, Kanai, et al., 1993).

Machine translation, document filtering, and other applications require a different measure of accuracy. Many new application specific metrics are needed to objectively assess progress made in OCR research. Examples of metrics and needed metrics are described in Rice, Kanai, et al. (1994); Kanai, Rice, et al. (1993).

Since a variety of factors affects the performance of OCR systems, a large amount of input test data must be used in the evaluation processes. Consider testing recognition of text printed in a variety of fonts. Over 3,000 combinations of typefaces and type styles are available for laser printers. If ten type sizes are used, over 30,000 test samples are required just to examine one instance of output for each input. Thus, automating both the measurement tasks and the analysis of data are essential. Aside from eliminating human error, automated experiments have the following benefits:

- Experiments are reproducible.

- The inherent consistency of automated systems tends to avoid bias toward algorithms by *excusing* certain types of errors.
- Large (statistically-significant) experiments can be conducted with little additional effort.

However, setting up automated testing systems (and metrics) is both costly and technically challenging. An example of an automated testing environment is described in Rice (1993).

There are different ways to prepare test data. Example sets of real-world document images with the associated *truth* representation are an ideal form of input test data. The *truth* representation and attributes of the input images must be manually prepared. Our experience shows that, it takes an average of 2 man-hours to prepare basic page-based data from a page, including the almost 100% accurate *truth* representation. Therefore, such data are extremely expensive.

It is also possible to generate simulated data. It is customary to perturb ideal images or sample hand-written characters by adding noise. Examples of distortion models are given in Ishii (1983); Baird (1992); Kanungo, Haralick, et al. (1993). This approach eliminates expensive *truth* preparation and allows researchers to control individual noise variables.

In spite of the appeal of generating large test databases this way, their value in predicting the behavior of OCR systems in field condition has not been established. The evaluation and comparison of real-world distortion (example sets) and simulated distortion are important new research tasks. Validation methods have been proposed by Nagy (1994); Li, Lopresti, et al. (1994).

Currently, most of the available databases are character-based. The ETL Character Database[†] mainly contains hand-printed segmented Japanese characters. The U.S. Postal Service[†] released a database containing hand-written characters extracted from envelope address blocks. The National Institute of Standards and Technology (NIST)[†] distributes a large number of hand-written segmented characters and hand-printed segmented characters.

The University of Washington[†] has released a database (UW-I) that contains 1,147 page images from scientific and technical journals with the corresponding *truth* representation. It also includes image degradation models and performance evaluation tools. The UW-II data set contains 43 complete articles in English and other data.

To objectively measure progress in character recognition technology and to identify research problems, two kinds of evaluation are needed: internal evaluation and

[†]See section 12.6.3 for contact addresses.

independent evaluation. In internal evaluation, researchers' own test data sets or standard (public) test databases are used to measure and compare their progress. The creation and distribution of a variety of standard test databases is an important task in the OCR research community.

Since character recognition systems can be customized or trained to accurately recognize a given set of data, independent evaluation is also required for objective final assessment. In independent evaluation, test databases are hidden from the development process.

In 1991, the Chinese government evaluated Chinese OCR systems developed under the State Plan 863 (CCW, 1991). Tests were strictly conducted using standardized data sets. The best machine-printed character recognition rates with and without context were 97.84% and 97.80%, respectively. The best hand-written character recognition rate without adapting to a particular user was 80%.

In 1992, the U.S. Census Bureau and NIST determined the state of the art in recognition of hand-written segmented characters (Wilkinson, Geist, et al., 1992). Twenty-six organizations from North America and Europe participated in this test program. About half of the systems correctly recognized over 95% of the digits, over 90% of the upper-case letters, and over 80% of the lower-case letters in the tests.

In 1992, the Institute for Posts and Telecommunications Policy in Japan evaluated OCR technology for recognizing postal codes (Matsui, Noumi, et al., 1993). Hand-written segmented character images were used to test systems. Five universities and eight OCR vendors submitted their systems. The highest recognition rate was 96.22% with the substitution error rate 0.37%.

Since 1992, the Information Science Research Institute at the University of Nevada, Las Vegas, has been conducting evaluation of OCR technology for recognition of machine-printed documents. In the 1994 study, six pre-release systems developed by commercial OCR vendors were tested using two sets of page images (Rice, Kanai, et al., 1994). These systems correctly recognized over 99% of the characters in good quality pages. However, there is a significant reduction in accuracy on poor quality pages. This study also includes other metrics, such as word accuracy, non-stopword accuracy, and automatic page segmentation.

Future Directions

In this rapidly evolving information age, the need for automated data entry systems is essential. To expedite progress in this field, there is a need for large quantities of both test and training data. This situation is likely to continue until the resources needed to provide such data are made available.

13.11 Chapter References

AMTA (1992). *MT evaluation: basis for future directions*, Washington, D.C. Association for Machine Translation in the Americas.

Arnold, D. et al. (1994). *Machine translation: an introductory guide.* NCC/Blackwell, Manchester, Oxford.

Arnold, E. et al. (1993). Special issue on evaluation of MT systems. *Machine Translation*, 8(1-2):1–126.

ARPA (1993a). *Proceedings of the 1993 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

ARPA (1993b). *Proceedings of the Fifth Message Understanding Conference*, Baltimore, Maryland. Morgan Kaufmann.

ARPA (1994). *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Baird, H. S. (1992). Document image defect models. In Baird, H. S., Bunke, H., and Yamamoto, K., editors, *Structured Document Analysis*, pages 1–16. Springer-Verlag.

Balkan, L. et al. (1994). Test suites for natural language processing. *Translating and the Computer*, 16:51–58. papers presented at a conference.

Bimbot, F. et al. (1994). Assessment methodology for speaker identification and verification systems: an overview. Technical Report SAM-A Project 6819, Task 2500, SAM-A, Martigny, Switzerland.

Black, E. (1993). Parsing english by computer: The state of the art. In *Proceedings of the 1993 International Symposium on Spoken Dialogue*, Waseda University, Tokyo.

Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Black, E., Garside, R., and Leech, G., editors (1993). *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach.* Rodopi, Amsterdam, Atlanta.

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

CCW (1991). Research achievements on Chinese character and voice recognition. *China Computer World*, 349. Written in Chinese.

Chollet, G., Bimbot, F., and Paoloni, A., editors (1994). *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland. ESCA.

Chollet, G., Capman, F., and Daoud, J. F. A. (1991). On the evaluation of recognizers—statistical validity of the tests. Technical Report SAM-ENST-02, SAM.

Cohen, P. R. and Oviatt, S. L. (1994). The role of voice in human-machine communication. In Roe, D. B. and Wilpon, J., editors, *Voice Communication Between Humans and Machines*, pages 34–75. National Academy of Sciences Press, Washington, DC.

Cole, R. A., Hirschman, L., Atlas, L., Beckman, M., Bierman, A., Bush, M., Cohen, J., Garcia, O., Hanson, B., Hermansky, H., Levinson, S., McKeown, K., Morgan, N., Novick, D., Ostendorf, M., Oviatt, S., Price, P., Silverman, H., Spitz, J., Waibel, A., Weinstein, C., Zahorian, S., and Zue, V. (1995). The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21.

Council, N. R. (1966). Appendices 9–15. In *Languages and Machines: Computers in Translation and Linguistics*. National Academy of Sciences, Washington, DC.

DARPA (1989). *Proceedings of the Second DARPA Speech and Natural Language Workshop*, Cape Cod, Massachusetts. Defense Advanced Research Projects Agency.

DARPA (1990). *Proceedings of the Third DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991a). *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1991b). *Proceedings of the Third Message Understanding Conference*, San Diego, California. Morgan Kaufmann.

DARPA (1992a). *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*. Defense Advanced Research Projects Agency, Morgan Kaufmann.

DARPA (1992b). *Proceedings of the Fourth Message Understanding Conference*, McLean, Virginia. Morgan Kaufmann.

Eagles (1995). Report of the spoken language systems working group 5. Technical report, EAGLES, EAGLES Secretariat, Istituto di Linguistica Computazionale, Via della Faggiola 32, Pisa, Italy 56126, Fax: +39 50 589055, E-mail: ceditor@tnos.ilc.pi.cnr.it. In press.

Eurospeech (1993). *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin. European Speech Communication Association.

Falkedal, K., editor (1994). *Proceedings of the of the Evaluators' Forum, 1991*, Les Rasses, Vaud, Switzerland. ISSCO, Geneva.

Fourcin, A. et al. (1992). ESPRIT project 2589 (SAM) multi-lingual speech input/output assessment, methodology and standardization. Technical Report SAM-UCL-G004, SAM.

Galliers, J. R. and Sparck Jones, K. (1993). Evaluating natural language processing systems. Technical Report 291, University of Cambridge Computer Laboratory. To appear in *Springer Lecture Notes in Artificial Intelligence*.

Harman, D., editor (1993). *National Institute of Standards and Technology Special Publication No. 500-207 on the The First Text REtrieval Conference (TREC-1)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harman, D. (1993). Overview of the first Text REtrieval Conference (TREC-1). In Harman, D., editor, *National Institute of Standards and Technology Special Publication No. 500-207 on the The First Text REtrieval Conference (TREC-1)*, pages 1–20, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harman, D., editor (1994). *National Institute of Standards and Technology Special Publication No. 500-215 on the The Second Text REtrieval Conference (TREC-2)*, Washington, DC. National Institute of Standards and Technology, U.S. Department of Commerce, U.S. Government Printing Office.

Harrison, P., Abney, S., Black, E., Flickenger, D., Gdaniec, C., Grishman, R., Hindle, D., Ingria, R., Marcus, M., Santorini, B., and Strzalkowski, T. (1991). Evaluating syntax performance of parser/grammars of English. In *Proceedings of the Workshop On Evaluating Natural Language Processing Systems*. Association For Computational Linguistics.

Hausser, R. (1994). The coordinator's final report on the first Morpholympics. *LDV-Forum*, 11(1):54–64.

Höge, M., Hohmann, A., and Mayer, R. (1992). Evaluations of TWB: Operationalization and test results. Final Report of the ESPRIT I Project 2315 Translators' Workbench (TWB).

Höge, M., Hohmann, A., van der Horst, K., Evans, S., and Caeyers, H. (1993). User participation in the TWB II project: The first test cycle. Report of the Esprit II Project 6005 Translators' Workbench II (TWB II).

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37:158–166.

Houtgast, T. and Steeneken, H. J. M. (1984). A multi-lingual evaluation of the Rasti-method for estimating speech intelligibility in auditoria. *Acustica*, 54:185–199.

Hutchins, W. J. and Somers, H. L. (1992). An introduction to machine translation. In *An introduction to Machine Translation*. Academic Press, London.

ICDAR (1993). *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Ishii, K. (1983). Generation of distored charaters and its applications. *System, Computer, Controls*, 14(6):1270–1277.

ISO (1991). Information technology—software product evaluation, quality characteristics and guidelines for their use. Technical Report 9126, International Organization for Standardization.

ITU (1993). ITU-TTS draft recommendation p.8s: Subjective performance assessment of the quality of speech voice output devices. Technical Report COM 12-6-E, International Telecommunication Union.

Jekosch, U. (1993). Speech quality assessment and evaluation. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 2, pages 1387–1394, Berlin. European Speech Communication Association. Keynote address.

Jones, K. and Mariani, J., editors (1992). *Proceedings of the 1992 Workshop of the International Committee on Speech Databases and I/O Systems Assessment.* COCOSDA.

Kanai, J., Rice, S. V., Nartker, T. A., and Nagy, G. (1993). Performance metrics for document understanding systems. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 424–427, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Kanungo, T., Haralick, R. M., and Phillips, I. (1993). Global and local document degradation models. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 730–736, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Karis, D. and Dobroth, K. M. (1991). Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal of Selected Areas in Communications*, 9(4):574–585.

King, M. and Falkedal, K. (1990). Using test suites in evaluation of MT systems. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 211–216, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Klaus, H., Klix, H., Sotscheck, J., and Fellbaumn, K. (1993). An evaluation system for ascertaining the quality of synthetic speech based on subjective category rating tests. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 3, pages 1679–1682, Berlin. European Speech Communication Association.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *J. of the Acoustical Society of America*, 34:1689–1697.

Lehrberger, J. and Bourbeau, L. (1988). *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins, Amsterdam, Philadelphia.

Li, Y., Lopresti, D., and Tomkins, A. (1994). Validation of document image defect models for optical character recognition. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 137–150, University of Nevada, Las Vegas.

Logan, J. S., Greene, B. G., and Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America*, 86(2):566–581.

Matsui, T., Noumi, T., Yamashita, I., Watanabe, T., and Yoshimuro, M. (1993). State of the art of handwritten numeral recognition in Japan—the results of the first

IPTP character recognition competition. In *Proceedings of the Second International Conference on Document Analysis and Recognition*, pages 391–396, Tsukuba Science City, Japan. AIPR-IEEE, IAPR.

Moore, R. C. (1994). Semantic evaluation for spoken-language systems. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Nagy, G. (1994). Validation of simulated OCR data sets. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 127–135, University of Nevada, Las Vegas.

Nerbonne, J., Netter, K., Diagne, A. K., Klein, J., and Dickmann, L. (1993). A diagnostic tool for German syntax. *Machine Translation*, 8:85–107.

Nomura, H. and Isahara, H. (1992). JEIDA's criteria on machine translation evaluation. In *Proceedings of the International Symposium on Natural Language Understanding and AI*, Kyushu Institute of Technology, Iizuka, Japan. part of the International Symposia on Information Sciences.

Oviatt, S. L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9:19–35.

Oviatt, S. L., Cohen, P. R., and Wang, M. Q. (1994). Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication*, 15(3–4):283–300.

Pallett, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., and Prysbocki, M. (1994). 1993 benchmark tests for the ARPA spoken language program. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, pages 49–74, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Pols, L. C. W. (1991). Quality assessment of text-to-speech synthesis-by-rule. In Furui, S. and Sondhi, M. M., editors, *Advances in speech signal processing*, chapter 13, pages 387–416. Marcel Dekker, New York.

Pols, L. C. W. (1994a). Speech technology systems: Performance and evaluation. In Asher, R. E., editor, *The Encyclopedia of Language and Linguistics*, volume 8, pages 4289–4296. Pergamon Press, Oxford.

Pols, L. C. W. (1994b). Voice quality of synthetic speech: Representation and evaluation. In *Proceedings of the 1994 International Conference on Spoken Language Processing*, volume 3, pages 1443–1446, Yokohama, Japan.

Pols, L. C. W. and Jekosch, U. (1994). A structured way of looking at the performance of text-to-speech systems. In *Proceedings, ESCA/IEEE Synthesis Workshop*, pages 203–206, New Paltz, New York.

Pols, L. C. W. and SAM-partners (1992). Multi-lingual synthesis evaluation methods. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, volume 1, pages 181–184, Banff, Alberta, Canada. University of Alberta.

Rhyne, J. R. and Wolf, C. G. (1993). Recognition-based user interfaces. In Hartson, H. R. and Hix, D., editors, *Advances in Human-Computer Interaction*, volume 4, chapter 7, pages 191–250. Ablex Publishing Corp, Norwood, New Jersey.

Rice, S. V. (1993). The OCR experimental environment, version 3. Technical Report ISRI TR-93-04, University of Nevada, Las Vegas, Nevada.

Rice, S. V., Kanai, J., and Nartker, T. A. (1993). An evaluation of OCR accuracy. Technical Report ISRI TR-93-01, University of Nevada, Las Vegas, Nevada.

Rice, S. V., Kanai, J., and Nartker, T. A. (1994). The third annual test of OCR accuracy. Technical Report ISRI TR-94-03, University of Nevada, Las Vegas, Nevada.

Rinsche, A. (1993). Evaluationsverfahren für maschinelle übersetzunngssysteme: zur methodik und experimentellen praxis. Technical report, Kommission der Europaeischen Gemeinschaften, Bericht EUR 14766 DE.

Sinaiko, H. W. and Klare, G. R. (1972). Further experiments in language translation: readability of computer translations. *ITL*, 15:1–29.

Sinaiko, H. W. and Klare, G. R. (1973). Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL*, 19:29–52.

Sorin, C. (1994). Towards high-quality multilingual text-to-speech. In *Proceedings of the CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology*, pages 53–62, Münich.

Sparck Jones, K. (1994). Towards better NLP system evaluation. In *Proceedings of the 1994 ARPA Human Language Technology Workshop*, Princeton, New Jersey. Advanced Research Projects Agency, Morgan Kaufmann.

Spiegel, M. F. (1993). Using the ORATOR synthesizer for a public reverse-directory service: Design, lessons, and recommendations. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 3, pages 1897–1900, Berlin. European Speech Communication Association.

Spitz, J. (1991). Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, Pacific Grove, California. Defense Advanced Research Projects Agency, Morgan Kaufmann.

Steeneken, H. J. M. (1992). Quality evaluation of speech processing systems. In Ince, N., editor, *Digital Speech Coding: Speech coding, Synthesis and Recognition*, chapter 5, pages 127–160. Kluwer Norwell, USA.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *J. Acoustical Society of America*, 67:318–326.

Steeneken, H. J. M., Verhave, J., and Houtgast, T. (1993). Objective assessment of speech communication systems; introduction of a software based procedure. In *Eurospeech '93, Proceedings of the Third European Conference on Speech Communication and Technology*, volume 1, pages 203–206, Berlin. European Speech Communication Association.

Thompson, H., editor (1992). *The Strategic Role of Evaluation in Natural Language Processing and Speech Technology.* Human Communication Research Centre, University of Edinburgh.

Van Slype, G. (1982). Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua*, 1(4):221–237.

White, J. S. et al. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 193–205, Washington, DC. Association for Machine Translation in the Americas.

Wilkinson, R. A., Geist, J., Janet, S., Grother, P. J., Burges, C. J. C., Creecy, R., Hammond, B., Hull, J. J., Larsen, N. J., Vogl, T. P., and Wilson, C. L. (1992). The first census optical character recognition systems conference. Technical Report NISTIR-4912, National Institute of Standards and Technology, U.S. Department of Commerce.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies*, 34:527–547.