Methods and Metrics for the Evaluation of Dictation Systems: A Case Study.

Maria Canelli, Daniele Grasso, Margaret King.

TIM/ISSCO, ETI

University of Geneva, 40 blvd du Pont d'Arve, CH-1211 Geneva 4. Margaret.King©issco.unige.ch

Abstract

This paper describes the practical evaluation of two commercial dictation systems in order to assess the potential usefulness of such technology in the specific context of a translation service translating legal text into Italian. The service suffers at times from heavy workload, lengthy documents and short deadlines. Use of dictation systems accepting continuous speech might improve productivity at these times. Design and execution of the evaluation followed the methodology worked out by the EAGLES Evaluation Working Group. The evaluation therefore also constitutes a test bed application of this methodology.

1. Background

This paper presents the application of the ISO/EAGLES methodology for the construction of evaluations (Bevan, 1997; EAGLES, 1996, 1999; King and Maegaard, 1998) to the practical evaluation of two commercial dictation systems. For the purposes of this paper, the two systems will remain anonymous.

2. The ISO/EAGLES Methodology.

The methodology worked out by the EAGLES Evaluation Working Group involves following a series of steps in designing and executing an evaluation. These steps are reflected in the headings of the remaining sections of this paper. Further detail and discussion can be found in the ISO and EAGLES references cited above. A summary accompanied by an informal worked example can also be found in King (1999).

The methodology proposed is generic, in the sense that it is set out in abstract terms which are meant to be applicable to the design of an evaluation for any type of human language technology product or system. The generic methodology will not be further discussed here, where we concentrate on its application to the design and execution of an evaluation of specific products carried out in the light of a particular intended context of use for those products.

3. Designing the Evaluation.

3.1 Context and purpose of the evaluation (task model).

The context of the evaluation was the daily work of a medium-sized governmental translation service. A major part of the translators' work is the translation of legal texts into Italian. Workload fluctuates severely, strongly influenced by external factors such as whether parliament is in session or not. All translators work directly with a text processor, although typing ability varies significantly. Translators are sometimes called upon to work very long hours, in situations where sometimes lengthy texts have to be treated with great dispatch. Heavy work loads, short deadlines and lengthy texts combine to produce a level of fatigue which contributes to a high error rate in typing, especially amongst those with weaker typing skills. The purpose of this evaluation therefore was to establish whether the use of a dictation system (i.e. a system which from spoken Italian input would produce an electronic version of the translation) might increase productivity by reducing the amount of time required to produce a correct electronic version of the translation.

There was no strong argument against the use of dictation systems coming from the work context itself: none of the translators suffers from any kind of speech defect or shows particular oddities of pronunciation, the working environment is quiet. Furthermore, the translators are not pre-disposed against the use of dictation systems in their work: some are enthusiastic about the idea, some interested but reserved, some neutral.

It should be stressed that the purpose of the evaluation was not to decide on the purchase of one of the two systems tested. Rather, the evaluation should be seen as part of a preliminary investigation into the potential usefulness of dictation systems in general. Thus, comparative evaluation of the two systems tested was only of subsidiary interest. This is of especial importance in a context where technology is evolving very rapidly: new versions of each of the products have subsequently appeared, and quite different comparative results might well be obtained if the later versions were to be tested. (It is also partly for this reason that the products are kept anonymous in this paper). Strenuous efforts were made however to ensure that the metrics developed for this evaluation could be validly applied to testing both newer versions and other commercial products.

3.2 Choice of products to be tested.

It was feared that testing only one product might produce results which, although valid for that particular product, did not represent the state of the art in speech technology in general: in other words, we might accidentally have hit on a particularly bad dictation system. On the other hand, limited resources meant that no more than two products could be tested. Obviously, any candidate product had to accept Italian input. The commercial systems finally tested were chosen mainly on the grounds that they seemed to be amongst the most widely used, and both had received favourable notice in the press and on the net. In the rest of this paper they are known simply as System A and System B.

The choice was confirmed by checking configuration requirements in order to ensure that suitable configurations meeting the product's minimum requirements were available both in the end-user environment and in the testing environment. We also checked that the configuration available at the testing site was comparable to that available at the end-user site, in order to avoid the risk that results obtained at the testing site could not be reproduced at the end user site.

3.3 Development of a quality model.

The EAGLES/ISO methodology requires the working out of a quality model, which specifies those attributes of the object to be evaluated which contribute to an assessment of its potential within a particular context.

ISO/IEC 9126-1 offers a starting point for constructing this model in the form of a hierarchically organized list of quality characteristics. These are best seen as a checklist helping the evaluation designer to structure his thinking rather than as a would-be exhaustive list where every item is to be taken into consideration in every evaluation.

ISO 9126-1 distinguishes between two types of software quality characteristics. Internal characteristics relate to attributes of the software itself - the algorithms used, the coding and so on. External characteristics concern what the user of the software actually sees. In this evaluation, internal characteristics are not relevant (or available), and the focus is on external characteristics. Thus, in terms used in other evaluations, the evaluation is black box rather than glass box.

The first major quality characteristic proposed is **functionality**, divided into five sub-characteristics: suitability, accuracy, interoperability, security and compliance. Of these five, we chose to concentrate on *accuracy*, defined by ISO as "The capability of the software product to provide the right or agreed results or effects" and interpreted by us in this particular case as the software's ability effectively to produce accurate written text on the basis of spoken input. We also chose to look at *interoperability*, defined by ISO as "The capability of the software product to interact with one or more specified systems." Here the main issue in our context was interoperability with the text processor used by the translation service end-users.

The second main ISO quality characteristic is **reliability**, defined as "The capability of the software product to maintain a specified level of performance when used under specified conditions", and further broken down into maturity, fault tolerance, recoverability and compliance. Of these sub-characteristics, we picked out *recoverability* as being of especial interest in our context. The definition given by ISO is "The capability of the software product to re-establish a specified level of performance and recover the data directly affected in the case of failure" which we interpreted to mean in our case the ability of the software to correct transcription errors on the basis of spoken corrections.

Usability, the third ISO quality characteristic, is defined as "The capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions." From its subcharacteristics of understandability, learnability, operability, attractiveness and compliance, we picked out *learnability* ("The capability of the software product to enable the user to learn its application") and *operability* ("The capability of the software product to enable the user to learn its application") and *operability* ("The capability of the software product to enable the user to operate and control it") as being pertinent to this evaluation.

The next major ISO characteristic, **efficiency**, we decided to ignore. This decision is a direct consequence of the purpose of this evaluation - trying to determine whether a technology is potentially useful in the specified work context. We felt that the application itself (recognition of continuous speech) guaranteed at least acceptable time and resource behaviour, and that, at this stage, we were not really concerned with any more detailed investigation of efficiency. Clearly, if the evaluation had been aimed at deciding whether to purchase a particular product, efficiency would assume a much greater relevance.

Maintainability is defined by ISO as "The capability of the software to be modified. Modifications may include corrections, improvements or adaptations of the software to changes in environment, and in requirements and functional specifications." Of its sub-characteristics, analysability, changeability, suitability, testability and compliance, we chose to focus on *changeability*, defined by ISO as "The capability of the software product to enable a specified modification to be implemented" and interpreted by us to mean in our specific case the ability of the software to learn from its user over time in order to improve the recognition of that user's voice.

Finally, we chose to ignore the sixth ISO characteristic, **portability**. This again is a characteristic which would become relevant if the evaluation were aimed at deciding whether or not to purchase the product(s) being tested, for example.

A third class of quality characteristics concerns what recent ISO proposals call "**quality in use**": this is quality as perceived by the end user, and reflects the combined effect of the internal and external software quality characteristics. It is divided by ISO into four subcharacteristics, effectiveness, productivity, safety and satisfaction. Of these, we chose to focus on *effectiveness*, defined as "The capability of the software product to enable users to achieve specified goals with accuracy and completeness in a specified context of use" and *satisfaction*, defined as "The capability of the software product to satisfy users in a specified context of use".

To summarize, the evaluation was to concentrate on:

External software	quanty characteristics.
Functionality	accuracy
	interoperability
Reliability	recoverability
Usability	learnability
	operability
Maintainability	changeability

External software quality characteristics:

Quality in use characteristics: effectiveness, satisfaction.

However, the quality characteristics chosen are not of equal importance. For example, there is little point in being able to learn how to use a product easily if the product gives totally unacceptable results. The next step, therefore, is to define the relative importance of each of the characteristics. It should be emphasized that this operation is very much specific to the particular evaluation to be done, its purpose and its context. Just as some quality characteristics become relevant when the purpose of the evaluation changes, the relative weight given to a specific quality characteristic may change drastically from one evaluation to another.

The table below summarizes in terms of points out of a hundred the relative importance we accorded to each characteristic. The quality in use characteristics are taken to be the sum of a number of external quality characteristics, and are inserted in the table at a point indicating what characteristics have contributed to them.

Functionality	Accuracy	60
-	Interoperability	5
Reliability	Recoverability	15
Maintainability	Changeability	15
Effectiveness		95%
Usability	Learnability	2.5
	Operability	2.5
Satisfaction		100%

As can be seen, accuracy has been taken to be far more important than anything else. Relatively low weights have been given to learnability and operability for three reasons. First, as will become clear below, in this evaluation these two attributes are judged subjectively, and subjective judgements are notoriously unreliable. Secondly, there are remedies for poor scores on these two attributes, in the form of improved documentation, training courses, help desks, advice from colleagues and the like. Thirdly, this particular evaluation was more concerned with the *potential* of dictation system technology to help with achieving translation tasks in a specific context, and less concerned with the usability characteristics of specific products.

Interoperability has received a low relative weight mainly at the end-users' request. They were reluctant to let factors which they saw as essentially implementation issues have too much effect on an overall assessment of whether a technology was potentially useful.

3.4 Developing metrics.

In this section we take each of the quality characteristics in turn and describe the metric associated with it.

Each metric will, when applied to the product being evaluated, produce a score. A three element rating scale is used to indicate whether that score is to count as a good, an acceptable, or an unacceptable score. The number of points awarded to the product is determined by its position on the three point scale. If the score is good, the total number of points is given, if acceptable, half of the available points, if poor, none at all.

An example may make this clearer. According to the relative weightings specified above, a total of 60 points may be awarded for accuracy. If a product gets a good score on the associated metric for accuracy, it will collect 60 points. For an acceptable score it will collect 30 points, for a poor score, it will collect none at all. When all the points awarded are added together, the product will have an overall score. The maximum score possible is 100. For each metric, we give also the rating scale.

Accuracy: the metric here is based on the ratio between the number of words wrongly transcribed and the number of words in the text. (This is very like the word transcription error rate metric widely used in the evaluation of speech recognition systems). Counting errors raises some issues which are addressed below in the section on executing the evaluation.

The rating scale for this metric was established by considering the number of typing mistakes typically produced by a tired translator. On this basis, the acceptability threshold was fixed at one error in every four words. This gives us the following rating scale:

Rating	Definition	Points
Good	Less than 1 error	60
	every 4 words	
Acceptable	1 error every 4	30
	words	
Poor	More than 1 error	0
	every 4 words	

In actual practice, the figure resulting from the calculation was converted to a percentage error rate by multiplying by 100. One error in every four words corresponds to a 25 % error rate. The table above can then be restated as:

Rating	Definition	Points
Good	< 25% error rate	60
Acceptable	25% error rate	30
Poor	> 25% error rate	0

Interoperability: the translators use Microsoft Word as their normal document production tool. For certain documents they use Excel, and they work, of course, with a specific operating system. However, the introduction of translation technology may imply the purchase of new computing machinery, so the operating system may well change. Word is used more widely than Excel. The metric therefore is a check on interoperability with Word, Excel and the local operating system.

The rating scale reflects the relative importance of these three pieces of software within the translation service. Interoperability with Word is regarded as essential. Interoperability with either or both of Excel and the local operating system is regarded as a significant advantage, but not essential. This gives us the following rating scale:

Rating	Definition	Points
Good	Compatibility with Word and	5
	Excel and/or the local operating	
	system	
Acceptable	Compatibility with Word only	2.5
Poor	Incompatibility with Word	0

Recoverability: this characteristic concerns the software's ability to correct errors on being given vocal instructions to carry out the corrections. The metric chosen here is the ratio between the time required to carry out the corrections and the number of errors requiring correction, i.e.

time for correction number of errors to be corrected The rating scale for this metric is based on a comparison between the time required for an experienced Word user to correct the text using the mouse and the keyboard with the time required for the same user to carry out the same set of corrections using only his voice. Here too the number of errors to be corrected is taken into account, in order to provide a unit of measurement allowing the two dictation systems to be compared. (It would be an unexpected coincidence if each system produced the same number of transcription errors). It was decided that it should not take more than one and a half times as long to correct vocally than to correct with mouse and keyboard. This gives us the following rating scale:

Rating	Definition	Points
Good	Voice correction faster than	15
	keyboard correction	
Acceptable	Voice correction takes as long	7.5
_	as or up to one and a half times	
	as long as keyboard correction	
Poor	Voice correction takes more	0
	than one a half times as long as	
	keyboard correction	

Changeability: this characteristic was defined as the software's ability to learn from use over time by one user in order to improve its recognition rate, or, in other words, to improve through training with a specific voice and a specific vocabulary. We took as a metric here the ratio between the number of transcription errors found the first time a text was dictated and the number of errors found when the same text was dictated a second time. (In order to avoid problems of reliability, recorded text was used for both dictations). Once again, the two systems tested are unlikely to produce the same number of errors, so in order to provide a comparable measure of improvement, we used the following formula, which will provide a figure that represents improvement expressed as a percentage

(Error count dictation 1) - (Error count dictation 2) Error count dictation 1

As mentioned previously, counting of errors presents some interesting problems which will be discussed below in the section on executing the evaluation.

For this characteristic, it was difficult to find a strong justification for a rating scale. Thus, it was intuition which suggested to us that a 30% improvement would be an acceptable result, and a 50% improvement a good result. Experience may well lead to these figures being revised. In the meanwhile, we have the following rating scale

Rating	Definition	Points
Good	50% or more improvement	15
Acceptable	30% - 50% improvement	7.5
Poor	Under 30% improvement	0

Learnability and **operability**, the two remaining characteristics, are exceptions in the work reported here, in that it was decided not to try to devise quantifiable objective ways of testing a system on these characteristics, but to trust to subjective judgement by the testers. Conscious of the tendency towards unreliability thus introduced, we accorded the two characteristics rather low

relative weights. For the same reason, we did not define an acceptability threshold for them.

3.5 Designing execution of the evaluation and preparing test materials.

The next step is to produce a plan for how the evaluation should be executed. This section describes the different steps which formed part of that plan. The major activity of each step is indicated by the phrases in bold print.

A preliminary was to decide on a choice of testers. Practical considerations limited the number of testers to two, both native Italian speakers with no noticeable speech defects or regional dialect features in their speech. Both are young adults and one was male, the other female.

The first step was **familiarization with the products** to be tested. The familiarization phase included reading the technical literature provided with the product, experimentation with the commands to be used in operating the products, and, finally, a first dictation, using the training text provided by the manufacturer. The decision to use the training text provided rather than a text more closely resembling the work of the translation service was based on a desire to give each product its best chance: it was unlikely that the products in their initial state would be rich in the vocabulary typical of the texts treated by the translation service, and this could easily prejudice adversely system performance. Using the training text provided at least allowed training on the testers' voices.

The familiarization phase also served as a basis for the judgements on learnability and operability.

The next step was a **first real dictation**, this time of a text extracted from a document provided by the translation service and therefore reflecting the texts routinely treated by them.

Testing dictation software presents a problem of reliability: if a text is dictated twice, even by the same person each time, it is impossible to ensure that the conditions are exactly the same on both occasions, or even that the dictation itself will have the same characteristics: speed of dictation may vary, hesitations may occur at different times, ambient noise may differ both in level and in quality. However, it was clearly important that the same text should be dictated to each of the two products to be tested.

It was therefore decided to adopt the strategy, known from previous evaluations reported in the literature, of using recorded text. However, this strategy did not resolve all problems. Both systems to be tested allow the dictator to specify capitalization, but the commands used by the two systems are different. It was therefore decided to omit capitalization commands from the dictation. However, parentheses, colons and semi-colons, full stops and paragraph breaks were dictated.

This first recorded dictation was also used as part of the test material for measuring changeability. The same text would be dictated a second time in order to determine to what extent the performance of the two systems had improved after correction of the transcription errors found after the first dictation. Once the first dictation was complete, the transcribed texts were saved and errors counted to provide data for the accuracy metric.

The next step was **correction of the first dictation**. Here, a major problem of comparability presented itself: one of the two systems did not allow errors to be corrected through spoken commands. It was therefore decided to award this system a zero score for recoverability, but nonetheless to carry out the corrections through the system correction interface provided, since the system is supposed to improve itself as a result of corrections. For the other system, it was decided to proceed as originally planned, carrying out corrections first through spoken commands and subsequently carrying out the same corrections using the mouse and keyboard. Both correction sessions were timed.

The next step was **training the dictation systems to deal with the texts typical of the translation service**. For this, a second text, similar to the text used for the first dictation (in fact, both were extracted from the same document) was used. Once again, recorded text was used, and the two transcribed versions corrected as part of the training.

The final step was the **re-dictation of the first text**, which, it will be remembered, was recorded. Subsequently, transcription errors were counted.

4. Executing the Evaluation.

Here too we followed the recommendations of the relevant ISO/EAGLES methodology, which foresees three steps in the execution of the evaluation:

- take measures
- compare with the previously determined satisfaction ratings
- assess the results

However, in order to avoid some unnecessary repetition, we shall describe the first two steps together. Once again, we shall take each characteristic in turn.

4.1 Taking measures and comparing with rating scales.

Accuracy.

The metric here is based on the ratio between the number of transcription errors and the number of words in the dictated text. The text used for the first dictation contained 384 words. The calculation is therefore

(Error count/384) * 100

It is here that question of what to count as an error becomes critical. We adopted the following rules.

As a general rule, we counted one error for each word of the original text which was not correctly transcribed. This implies that even if the product incorrectly produces several words for one word in the original text, this still only counts as one transcription error. Thus, one of the two products transcribed the original word "durevolmente" by the two words "dure" and "mentre". This was counted as one error.

All abbreviations were considered as one word. Thus "ST", "AG" and "SDM" were each counted as one word, even if the abbreviation was dictated by dictating each letter separately.

We did not consider it a transcription error if the system wrote out alphabetically in full numbers which appeared in numerical characters in the text, or, inversely, numerical characters for alphabetic renderings of numbers. The reason here comes from a comparability problem; both products allow commands specifying whether numerical or alphabetic characters are to be used, but the commands are different. Recording the text meant that it was impossible to use both sets of commands. Since the problem presented itself rarely in the test text, we took the easiest option of simply omitting the relevant commands, and accepting either form of transcription as correct.

However, we did take as an error the decomposition of a number into its different elements, for example "250 otto" instead of "258".

Taking into account these rules for counting errors, the two products produced surprisingly disparate results. System A produced a total of 35 errors, where System B produced 103. Applying the formula specified to these raw figures we get the following results for accuracy:

System	Score	Points
А	< 25% error rate	60
В	> 25% error rate	0

Interoperability.

This measure was taken by consulting the technical documentation of each product and confirming positive responses by experimentation. Both systems were interoperable with Word. However, System A was far more limited in interoperability with softwares other than Word. This gives us the following results for interoperability:

System	Points
А	2.5
В	5

Recoverability.

The metric here, it will be remembered, is based on the time taken to correct the errors using spoken commands compared to the time required to carry out the same corrections using mouse and keyboard, divided by the number of transcription errors in order to obtain a comparable measure. It will also be remembered that one of the systems tested, System B, did not allow for correction through spoken commands. We therefore only have results here for System A, and the comparative calculation became irrelevant.

With System A, correction through spoken commands (T1) took 38 minutes, correction through mouse and keyboard (T2) took 11 minutes. The rating scale for this characteristic defines that the result is poor if T1 > 150%T2.

We thus get the following results for recoverability:

System	Points
А	0
В	0

It is perhaps worth reminding the reader that System B's errors were nonetheless corrected using the correction interface window offered by the software. This correction took quite a long time since System B automatically requests training for those words which do not appear in its initial vocabulary.

Changeability.

The metric here is based on the number of transcription errors when the same text is dictated twice, with correction of errors having been carried out between the first and second dictations:

(Error count dictation 1) - (Error count dictation 2) Error count dictation 1

The results here for System A were somewhat surprising: having made 35 transcription errors on the first dictation of the test text, System A made 44 transcription errors on the second dictation of the same text. This gives a negative score for changeability of -26%. System B, however, showed significant improvement. Having made 103 errors on the first dictation, 59 were made on the second dictation, giving a score of 43%. The rating scale specifies an improvement of between 30% and 50% to be an acceptable score, thus giving the following results for changeability:

System	Score	Points
А	-26%	0
В	43%	7.5

Learnability.

It will be remembered that points were given for this characteristic based only on subjective judgement by the testers. It seems nonetheless worth while to record some of the factors that influenced the points awarded.

A major negative factor with System B was that no user manual is supplied automatically with the product. The on-line help was printed out in order to carry out familiarization with the product, but this is obviously only a second rate substitute for proper documentation in the form of a user manual.

The testers were rather unhappy with the training texts and procedures offered with both systems. System A asked the user to dictate part of a rather well known children's story. Whilst this was quite fun, the text is hard to dictate, since it contains both archaic expressions and some phraseology which is close to regional dialect. Nor can the text be said to resemble the sort of text that the user is likely to use the system for. System B asked the user to dictate a series of sentences rich in economic vocabulary. Unfortunately, it was rather obvious that these sentences were often poor Italian translations of an English original, which is rather off-putting for users who are translators.

Finally, both our testers found the voice heard by the user during the training session to be quite disagreeable in the case of System A, quite pleasant in the case of System B.

The points finally awarded for learnability were:

System	Points
А	1.25
В	0

Operability.

There is very little to be said about operability: both systems use an interface which is very like the Word interface, and therefore very familiar both to the testers and to the translator end users. System B suffers from one minor defect: some words do not appear on the screen in their complete form - either the beginning or the end is missing. This is probably due to poor localisation: the translations into Italian have produced longer character strings than the corresponding English original, without this being accounted for.

The points awarded for operability are therefore:

System	Points
Α	2.5
В	1.25

4.2 Assessing the results.

The final step of the evaluation is to assemble the results and assess them. The tables below summarize the results for each system, where QC stands for Quality Characteristic, and where the maximum points available for each attribute (quality sub-characteristic) is shown in brackets after the number of points actually awarded:

QC	Attribute	Threshold	Points	
Functionality	Accuracy	1 error in 4 words (25%)	60 (60)	
	Interoperability	Compatibility with Word	2.5 (5)	
Reliability	Recoverability	Correction with spoken commands takes no more than one and a half times as long as conventional correction	0 (15)	
Maintainability	Changeability	30% - 50% improvement	0 (15)	
Effectiveness (sum of points to now)				
Usability	Learnability	Subjective	1.25 (2.5)	
	Operability	Subjective	2.5 (2.5)	
Satisfaction (total number of points awarded)				

It will be remembered that the main purpose of the evaluation was to discover whether use of a dictation system would help to improve productivity in a situation where tired translators tend to make a lot of typing mistakes. Looking at this summary allows us to give a positive response to this question, but with some reservations. Accuracy, with this system, is encouragingly high, but both recoverability and changeability are low. Recoverability reflects how easy it is to correct the transcribed text using spoken commands rather than mouse and keyboard. The results here would suggest that dictation followed by mouse and keyboard correction is the better option when productivity is the issue. Changeability reflects improvement over time, as a result of training through continued use by the same user. System A's performance actually deteriorated with respect to this attribute. However, the result should perhaps be treated with some caution, since only two dictation sessions contributed to the training. If evaluation were aimed at deciding whether or not to purchase this system, it would be important to carry out more exhaustive tests for this attribute.

System B

QC	Attribute	Threshold	Points	
Functionality	Accuracy	1 error in 4		
		words (25%)	30	
			(60)	
	Interoperability	Compatibility		
		with Word	5	
			(5)	
Reliability	Recoverability	Correction		
		with spoken		
		commands		
		takes no more		
		than one and		
		a nan umes		
		as long as		
		conventional	0	
		correction	(15)	
Maintainability	Changeability	30% - 50%	< - /	
5	6 5	improvement	7.5	
		1	(15)	
Effectiveness (sum of points to now)				
	_		(95)	
Usability	Learnability	Subjective	0	
			(2.5)	
	Operability	Subjective	1.25	
			(2.5)	
Satisfaction (total number of points awarded)				

The most immediate comment to make on the summary of results for System B is that the decision to test at least two dictation systems was obviously justified; if only System B had been tested, it would have been tempting to conclude that spoken language technology was not yet ripe enough to contribute to improved productivity.

5. Discussion and Conclusions.

Following the EAGLES/ISO methodology greatly facilitated defining an evaluation of specific products in a specific work environment. The only difficulty encountered was in separating out the elaboration of a task model from defining the purpose of the evaluation. In the EAGLES/ISO methodology these tasks are chronologically ordered, with the definition of the purpose coming first. Our difficulty was perhaps one of description in this paper; in order to describe the purpose to a reader not familiar with the translation service, we had to introduce description of that service, and thus start prematurely on the construction of a task model.

The work described here also confirms the advisability of interpreting the ISO quality characteristics as a check list and as an aid to structured thinking, as is suggested in the EAGLES reports. To take them as anything stronger, for example as a fully defined exhaustive list of features to be taken into consideration, leads to unnecessary and time wasting attempts to decide what universal and unequivocal meaning can be assigned to the names and definitions given in the ISO documents.

The primary weakness of the work reported here is that limited resources prevented the testing of more than two systems and the use of more than two testers, as well as curtailing testing of improvement after training. Using a greater range of speakers would improve the validity of the metrics. It is possible, too, that using only one type of text affects the validity of the metrics. An interesting experiment would be to repeat the evaluation exercise using a number of different kinds of texts in order to test this hypothesis.

With these reservations, we believe the metrics developed to be both valid and reliable, and thus potentially useful in other evaluations of dictation systems. An open question is how much they would carry over to the evaluation of spoken dialogue systems, where interaction with the user confuses certain issues.

6. Acknowledgements.

The authors would like to thank Véronqiue Sauron of the ETI for her unfailing help and support, and Dario Scazziga of the Swiss Federal Chancery for his invaluable cooperation.

The work was carried out on behalf of the Swisstra Association, whom we also thank.

7. References

- Bevan, N. (1997), ISO 9126. Presentation at the *EAGLES Evaluation Workshop*, Brussels, November 1997.
- Boisen, S. and Bates, M. (1983). A Practical methodology for the Evaluation of Spoken Language Systems. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*.
- EAGLES (1996) LRE-EAGLES Evaluation Working Group Final Report. Available from http://wwwissco.unige.ch/projects/eagles
- EAGLES (1999) LRE-EAGLES Evaluation Working Group Draft Final Report. Available from http://wwwissco.unige.ch/projects/eagles
- ISO (1991) ISO/IEC 9126 Quality Characteristics and Guidelines for their use. ISO, Geneva
- King, M. (1999) The 7-step Recipe for Evaluating Language Technology. In *Proceedings of the European Evaluation of Language Systems (EELS) Conference*, Hoevelaken, April 1999.
- King, M. and Maegaard, B. (1998). Issues in Natural language Systems Evaluation. In *Proceedings of The Language Resource and Evaluation Conference* (LREC), Granada, Spain, 1998.