[*Translating and the computer 13: The Theory and Practice of Machine Translation – A Marriage of Convenience?*
Papers presented at a conference … 28-29 November 1991 (London: Aslib, 1992)]

FROM EVALUATION TO SPECIFICATION

R Lee Humphreys

University of Essex

SUMMARY

The central and unresolved problem in evaluation is that of providing MT performance metrics which are meaningful to potential users. The traditional transformer architecture of older MT systems, and their use as general purpose machines, results in output text of rather poor quality which is difficult to grade. More recent architectures which use explicit (rule-based) representations of linguistic knowledge will tend to produce output of a consistently higher quality. Linguistic knowledge must be complete (relative to some intended coverage) and consistent. Hence, the increasing use of test suites as tools for evaluation against a specification. Designing MT systems for controlled language input is a further indicator of the growing significance of linguistic specification.

TRANSFORMER ARCHITECTURES AND THE TRIANGLE

Many, perhaps most, of the older commercial MT systems attempt to translate by performing extensive cosmetic surgery on the input text. For such systems, translating into French would be a matter of Frenchifying some English by changing the words and re-ordering them according to a few generalisations or prejudices about the differences between the two languages ("put adjectives *after* the noun ..."). These systems do not actually translate into French; how could they, given that the only thing they know about French is that it is something English might become if you nip and tuck it enough. Sometimes, by a happy coincidence, one of their output sentences might look like French; most of the time, however, the result is Quasi-French – a non-natural language whose structure and grammar is known to neither the system, nor its designer, and certainly not to the user.

When we come to evaluating the performance of such transformer systems, we are thus faced with the following unhappy triangle (Figure 1). English sentences are transformed into Q-French sentences. Although the structure of Q-French is completely unknown, it is assumed to have some intended relation to French. The system user's task is to guess from the Q-French sentence what the French sentence it failed to produce might be. And, in turn, the evaluator's task is to find out how difficult the user's task is.

Translational Quality Metrics -- Intelligibility

A traditional way of assessing the users task is to assign scores to Q sentences: top marks for those that look like perfect facsimiles of real language sentences and bottom marks for those that are so badly degraded as to prevent the average user/evaluator from guessing what a

[1] Or *a* French sentence, since we are dealing with a translation *relation.*

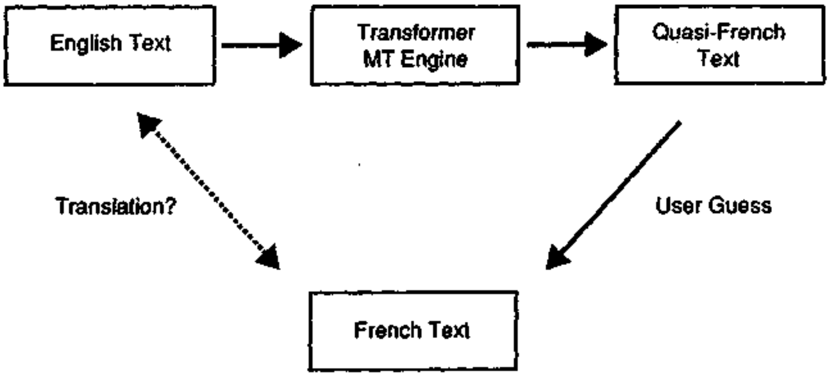

Figure 1 The unhappy triangle

reasonable real sentence in the context might be. In between these two extremes, Q sentences might be assigned higher or lower scores depending on their degree of awfulness. For example, minor gender or number agreement errors or slightly fluffed word order ("... *in an interview referred Major to the economic situation*") will probably get a better score than something where the main verb forms are not distinguished from infinitives, or with more-or-less uninterpretable lexical selection ("... *the peace contract should take off the peace agreement...*).

From a slightly different perspective, Q sentences can be seen as natural language sentences that have been degraded in some way; like a signal on a noisy line. Hence the relationship between the two. The dimension on which we are placing our scores was termed Intelligibility in the elegant and thorough study conducted by Carroll for the notorious ALPAC report [7]. For general discussions of quality metrics in MR see [3, 6, 9, 11].

Is there any principled way of constructing an intelligibility scoring system? Not that we know of. In a recent study of a small commercial MT system, the Evaluation Group with which the author is associated constructed simple scoring schemes which relied on informal descriptions and appropriate examples of typical deficiencies [2]. The wording and exemplification of the scoring instructions were progressively refined until some degree of scoring consistency was achieved within groups of otherwise untutored scorers working on the same material. In effect our scheme was designed to assess the global degree of intelligibility of Q sentences based on a handful of broadly characterised lexical and structural diagnostics or indicators.

In our evaluation we used a 1-4 scale for Intelligibility (rather than the 1-10 in Carroll's work). A larger scale entails that one can have a finer grained analysis of intelligibility variation between sentences; on the other hand, achieving statistically significant results with large scales will tend to require either analysis of a larger sample of Q sentences or an increase in the number of scorers. Both options add to expense.

Rather than using broad indicators as guides to score assignments, one could attempt a very precise typology of errors and deficits whose individual values are summed for each sentence to give an error/deficit score. Because some types of error/deficit are deemed more serious than others, it is usually thought desirable to weight scores in some way. There are two immediate problems with using detailed error analysis techniques of this sort. The first is practical: it will usually require considerable time and effort to train scorers to identify instances of particular errors and they will also need to spend more time analyzing each Q sentence. The second is more fundamental: for some MT systems, many Q sentences are so corrupted with respect to natural language correlates that detailed analysis of errors is not meaningful. Error types are not independent of each other: failure to supply any inflectional information for a main verb means that, necessarily, subject-verb agreement is undefined. It will be difficult to specify where one error starts and another ends and thus there is the risk of ending up with a general error scale of the form *one, two, ... lots.*

Accuracy

A Q sentence assigned a high intelligibility score is one for which it is very easy to guess a natural language correlate. However, it is always possible that this guess turns out to be quite inappropriate, i.e. it is not and could not (in the context) be a translation of the source sentence. Accuracy or Fidelity, the second important dimension of translation performance, is thus a relation between source sentences and putative translation in the target language which have been inferred by the user from Q sentences (the third side of the triangle).

Obviously enough, if the Q sentence is complete gobbledegook there is no known target language correlate and hence the notion of translational accuracy is simply inapplicable. If the Q sentence is intelligible in some degree then a target language correlate can be guessed and this may or may not prove accurate in some degree. As with Intelligibility, some sort of scoring scheme for Accuracy must be devised. Whilst it might initially seem tempting to just have simple Accurate and Inaccurate labels, this could be somewhat unfair to high performance MT systems where the inferred translation differs only in minor respects from the original, e.g., in the gender of a pronoun. Such a system would be deemed just as inaccurate as an automated Monty Python phrasebook which turns the innocent request *Please line my pockets with chamois*[2] into the target language statement *My hovercraft is full of eels.*

As it happens, in the sort of evaluation considered here Accuracy scores are much less interesting than intelligibility scores for the simple reason that if a Q sentence readily prompts a guess as to its natural language equivalent then most of the time for most systems that guess will turn out to be translationally accurate. That is, most systems most of the time do not exhibit Monty Python properties[3]. For some purposes it might be worthwhile to avoid

[2] I found this in an Italian-English phrasebook of the 1920s in the section on Talking to One's Tailor.

[3] Quite often one encounters Q sentences of the sort *The soldiers were in the coffee* (from, I think, *The soldiers were in the café)* but on some scoring schemes this might have a low intelligibility score anyway.

scoring for Accuracy and simply count up the number of cases where the inferred target language translation turns out to be misleading (leading one to suppose something which is actually incompatible with the source language sentence, although this relation would need extensive characterisation).

INTERPRETING QUALITY METRICS

It should be apparent from the above that devising and assigning quality scores for MT output, what we have called Declarative Evaluation, is not straightforward. Nor, however, is their interpretation.

MT is used either as a factor in the production of high-quality text translation (e.g., via post-editing) or as a means of allowing someone who has no knowledge of a particular source language to assess which parts of an incoming text stream in that language look sufficiently interesting or relevant to his/her interests to justify commissioning a high-quality translation (Gisting). It is virtually impossible, even for the evaluator, to decide what a set of Intelligibility and Accuracy (or whatever) scores for a single MT system might mean in terms of these end users.

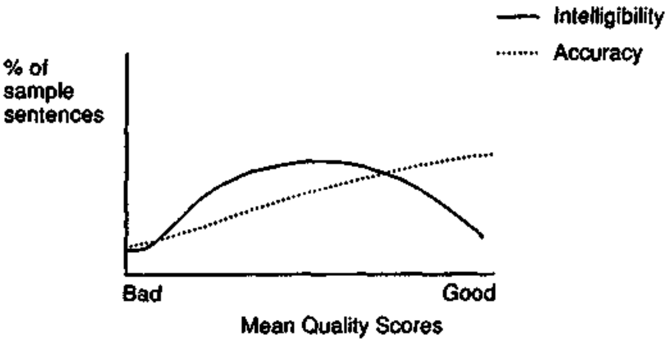

Figure 2 Typical quality profile for an MT system

If roughly half the Q sentences obtained from the evaluation corpus are of middling intelligibility, does that mean that you can use the system to successfully gist agricultural reports? One cannot say (and this is only in part the result of ignorance of the user's personal success criteria for gisting).

Turning to the high-quality translation case, it is clear that substantial post-editing will be required. But it is not clear, without further information about the relationship between measured quality and post-editing times, what effect on overall translator productivity the system will have. Whilst it is presumably true that increasingly unintelligible sentences will tend to be increasingly difficult to post-edit into shape, the relationship may not be linear. For example, it may be that sorting out minor agreement problems (which do not affect intelligibility very much) is just as much of an editing problem as sorting out major lexical selection problems (which affect intelligibility a great deal). Furthermore, translators may spend a lot of time trying to interpret and edit low-grade output when it might be quicker in many cases to translate from scratch.

It is true that a comparative evaluation of a number of different MT systems might demonstrate that one system is in all respects better than the others. However, practical knowledge of the relationship between editability and Intelligibility/Accuracy is still needed to assess whether this difference is sufficient to have any practical significance. If both systems are appallingly bad, it matters little that one is less appallingly bad than the other. And even if two systems have different performance profiles, it may not always be clear whether one profile is likely to be more matched to the task in hand than the other.

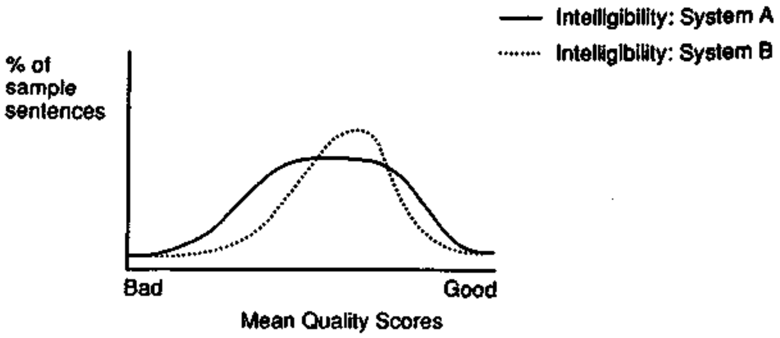

Figure 3 Quality scores for two MT systems

Another difficulty in interpreting the results of a declarative evaluation is that the test corpus may contain text types which are quite different from those of the prospective user. Performance on translating a series of news features is not a reliable guide to that which might be expected on translating scientific abstracts. Although one can imagine tools and techniques for automatically determining the syntactical and lexical similarity of different techniques, they are not yet widely available if they are available at all.

OPERATIONAL EVALUATION

The difficulties in designing and conducting an operation evaluation, and in interpreting its results, might seem to suggest that a different evaluation strategy is needed. Instead of trying to establish performance in absolute terms (Intelligibility/Fidelity scores, or error counts), one could try to directly assess the costs of producing translation of the required quality using the MT system as a production aid[4].

But this sort of evaluation, operational evaluation, will not be easy or cheap either. Suppose you are a potential user of some particular MT system which the vendor has lent you on a try before you buy basis. Once you have got the system up and running and figured out how to connect it into your overall document processing system, your translator team will have to be trained to cope with the post-editing environment. They will also require time to become familiar with the cognitively unfamiliar task of post-editing raw MT output. This is a skill which must be acquired and not all translators will acquire it as quickly or thoroughly as their peers.

There is a further problem. Most of the specialist terms used to describe your company products and services will not be known to the system. Naturally, they can be added in to the various system dictionaries, together with their translations, but that takes time and presupposes a certain training period. There will also be certain unusual types of multi-word construction and collocations which occur in your sort of text with high frequency; once the system knows about them, it will produce better quality translations. Learning how to tell the system about these constructions, and actually doing it, will also take some time. It could easily lake many translator-months before it starts to become apparent what sort of contribution the customised system could finally make to the translation task. Twelve translator-months (a not unrealistic assumption) spent on such an evaluation adds up to a lot of money. In some cases this could approach or even exceed the cost of the system itself. The advantage of operational evaluation is that the performance of the MT system is directly determined in terms of its cost contribution and hence no further interpretation is required. It might seem that it would be sensible for an MT supplier to carry out an operation evaluation and publish the results. But this information, although perhaps furthering the cause of MT in general, could not be readily interpreted by other users either. Quite apart from the continuing difficulty of text type, the model translation organisation in which the evaluation is conducted may be thought to be significantly different from the candidate user's own organisation in some respect or other and, hence, unrepresentative for that user.

LINGUISTIC KNOWLEDGE ARCHITECTURES AND THE TEST SUITE

In previous sections we have looked at the evaluation of commercial MT systems, with particular reference to those which work by transforming input strings. For some years the trend, at least in research circles, has been towards systems with substantial linguistic knowledge in the form of rule-based grammars. These systems have a different performance profile: typically, either they deliver a well-structured output sentence, albeit stylistically

[4] Vasconcellos' study [12] serves as an example although the author provides no final cost comparisons.

inappropriate, or they deliver nothing at all. Output will tend to contain rather fewer badly degraded sentences.

This behaviour reflects the fact that, at the core of such systems, either a structure is recognised by at least one linguistic rule at each conceptual level or it is not. If it is, the processes of informed analysis and synthesis can continue and if it is not, the system simply halts in an error state. In the older transformer architecture, rules have a much more permissive character. If a structure is recognised by a rule at some particular point in the process, that is fine; if it is not recognised by any rule at that particular point, that is fine too. In the very worst case, transformer architecture can return a completely unchanged input string. In the very best case, it can return something which looks astonishingly like a well-formed target language sentence. In the average case, it returns neither one thing nor the other.

Assessing the quality of transformer output, assigning quality scores, is difficult because one is being asked to compare ideal or acceptable sentences with strings which are not part of any language at all. By contrast linguistic-knowledge MT output strings at least belong to a formal language; a formal language which approximates to a human language. Errors and deficits with respect to acceptable natural language translations will tend to fall into better defined classes and hence quality assessment should become a manageable task[5].

Although the use of linguistic-knowledge based techniques tends to promote high Intelligibility and Accuracy output, it is always possible that the linguistic knowledge embedded in the system is simply wrong. Sometimes the computational linguist writing the rules for some part of the system fails to recognise that a particular rule admits or excludes more linguistic circumstances than anticipated; sometimes the rule set needed to handle a particular phenomenon is incomplete; rather frequently, a complete rule set handling phenomenon A, e.g., modal auxiliaries, and another rule set handling phenomenon B, e.g., negation, fail to work correctly together when the two phenomena co-occur or interact in a sentence.

Keeping track of these sorts of constructional errors and deficits has become rather a severe problem for developers of large natural language processing systems. For example, whilst running the system on a corpus of test texts will reveal many problems, many potential areas of difficulty are hidden because the statistics are such that even quite large corpora will lack even a single example of some particular grammatical combination of linguistic phenomena.

Rather than churning through increasingly large natural text corpora, the R&D community has recently turned its attention to the use of suites of specially constructed test sentences [4, 13]. Each sentence in the suite contains either one linguistic construction of interest or a combination thereof. Thus part of an English test suite might look as follows:

[5] Practical and commercial realities require that if a system based on linguistic knowledge is to be used for real work, its core engine must be supplemented with various other components, coping strategies, that enable it to perform gracefully when a word is not recognised or when an input sentence is only partly grammatical in terms of the existing system grammars. Where the system has recourse to its coping strategies, one would expect the output to be significantly degraded, perhaps to the level of the brain-dead Q sentences characteristic of transformer architectures.

John runs.
John has run. *aspectual auxiliary*
John will run. *modal auxiliaries*
John can run.
John may run.
John should run.
John will have run. *modal and aspectual auxiliaries*
John may have run.
John should have run.
John can have run.

John does not run. *negation (with do-support)*
John not run.
John has not run. *negation and aspectual auxiliaries*
John will not run. *negation and modal auxiliaries*
John may not run.
John should not run.
John could not run.
John will not have run. *negation, modals and aspectuals*

A test suite may include grammatically unacceptable sentences, e.g., *John not run,* which ought not to be translated. In systems which use the same linguistic knowledge for both analysis and synthesis, the fact that an ill formed sentence is rejected in analysis suggests that it is unlikely to be constructed in synthesis either.

It is not entirely clear how test suites, which are a general NLP development test tool, should be constructed and deployed in the specific context of MT. For a bi-directional system one wants to have test suites for both the source and target language. Thus success in translating all the sentences in a German test suite into English and all the sentences in an English test suite into German would definitely be encouraging. However, standard test suites are designed to probe possible failings in the treatment of a single language and, as such, they are rather blunt instruments for probing translation performance. For example, suppose one part of an English test suite is intended to address "tough-movement" constructions. It might contain the following sentences:

John is easy to convince.
This book is easy for me to force Harry to read. *unbounded*

Whilst the unbounded tough-movement happens to be relevant and interesting for translation into German (which allows only the bounded variety), these sentences do not probe limitations on lexical triggers for tough movement in German:

Linguistics is boring to study.
Die Linguistik ist langweilig zu studieren.
Linguistik zu studieren ist langweilig.

There is no particular reason why the English test corpus should include two examples of a raising adjective (easy, boring) unless it is specifically designed for probing German translation at the same time. Whilst the corresponding German test corpus may contain both *leicht* and *langweilig,* these will probably be embedded in their permissible environments and not others. Thus running the MT system in both directions on each of the two test suites will fail to reveal whether *boring* really is translated correctly into the non-raised German form.

Given this sort of problem it is assumed that monolingual test suites should be supplemented by further sentences in each language designed to probe specific language pair differences [3]. Presumably some help can be had here from the comparative grammar tradition.

It should be clear that results with test suites can only be indicative. Linguists make test suites by listing all the linguistic phenomena they can think of in increasingly elaborate combinations. Many combinations which they forget or choose not to include (to keep the suite within manageable proportions) will turn out to be just those that the system cannot handle correctly.

Evaluation vs Specification

We have suggested that test suites have an important role in the development of current MT and NLP systems. What relevance do they have for the candidate user?

It is perfectly possible for the user to run an MT system on a test suite of their own devising and, in some cases, this may be perfectly appropriate. However, apart from the difficulty of knowing how to design a test suite (this is still a research problem), and the cost of actually constructing it, there remains a familiar problem: how are the results to be interpreted? Suppose System A and System B both produce acceptable translations for 40% of the test sentences and that they actually fail on different, or only partially overlapping, sentence subsets. Which one is better? If System B, but not System A, fails on test sentences which embody phenomena with very low historical frequencies in the user's type of text materials, then clearly System B is the better choice. But users typically do not have reliable information on the relative frequencies of various types of construction in their corpora and this can only be obtained by using automatic tools and techniques which are not yet widely available[6].

We have seen throughout the discussion of evaluation that variation in text type remains a problem. If some sort of evaluation procedure can show that an MT system performs quite well on one sort of text, this cannot be taken as a reliable prognosis of satisfactory performance on another type. By the same token, if a system seems to perform rather poorly on one text type, the supplier can always claim that its performance will be better on another. But perhaps this text-type problem can be set aside by changing the way suppliers and users think about MT systems.

Under the old scheme of things, one evaluated a system to find out whether it translated at all. Since newer systems contain systematic representations of linguistic knowledge, it now makes sense to evaluate against a linguistic specification. Using test suites to provide coverage of linguistic phenomena is an example of that.

The notion of specification can be pushed further. At present, MT systems tend to be offered as general-purpose machines. Suppliers routinely warn against using it on poetry, but they do not in general offer a system which has been set up for one particular type of

[6] The same problem of interpretability suggests that there would also be little point in having MT systems evaluated by an independent agency using some sort of standard set of test suites.

translation task. One obvious track for suppliers to take would be to insist that input text must have a rigorously controlled syntax and vocabulary which is exactly consonant with the linguistic knowledge embedded in the system. In principle, the performance of controlled-input systems of this sort should be much better than that of general purpose ones because (a) developers would have a smaller and more precise rule set to debug and (b) it would rarely, if ever, be necessary to resort to coping strategies to deal with unanticipated linguistic structures. Even the use of controlled input with ordinary general purpose MT systems can produce dramatically improved performance [10].

It might be argued that to insist on controlled input is to merely replace extensive post-editing with extensive pre-editing. This would be true if documents to be translated were originated in free text. However, in many corporate contexts it is now quite normal to write technical documents, for example, in controlled subsets of English to improve readability for both native and non-native English speakers. In this circumstance no specific pre-editing for MT would be required. Moreover, controlled language input could improve the performance of other NLP tools that might be used in document processing and maintenance, e.g., intelligent indexing systems, intelligent abstracters and so on.

Of course, one can give authors various instructions which encourage them to write in a controlled way; but there is no guarantee that they will in fact do so. Controlled language MT can only be optimised if the author is provided with support systems (parsers, intelligent sentence-completion prompters, etc.) which use the same linguistic knowledge component as the MT engine, thereby ensuring that text is truly conformant to the controlled language specification.

CONCLUSION

There are two themes running throughout this paper. Firstly, it is very difficult to evaluate an MT system in such a way that the results are readily interpretable by candidate users. Quality scores cannot be interpreted without further knowledge of their relationship to editability; test suite scores cannot be interpreted without further knowledge of the text frequency of the phenomena in the test sentences. Whilst an operational evaluation may give one user good insight into a system's cost-effectiveness, the evaluation will itself be costly and it is unlikely that other users can use the results to reliably predict cost-effectiveness for their translation task.

Secondly, the expected performance profile of newer linguistic knowledge-based systems is starting the change the nature of the evaluation task. The very fact that such systems have linguistic knowledge allows developer and user to specify what that knowledge should be[7]. Restricting the task domain will allow a reduction in errors and the general exclusion of the brain-damaged output so characteristic of the older transformer architecture.

Although the interpretability problem, the first theme, remains unresolved, the fact that the output of the newer systems looks more like natural language should make it easier for everyone to see whether or not it might be useful.

[7] The desirability and importance of specification has also been emphasised by Steven Krauwer in an address given in April 1991 to the Evaluators Forum at Ste Croix, Switzerland.

REFE RENCES

[1] Douglas Arnold (1990), Text Typology and Machine Translation: an overview, *Translating and the Computer,* 10, Aslib, London.

[2] L. Balkan, M. Jaaschke, L. Humphreys, S. Meijer and A. Way (1991), Declarative Evaluation of an MT System: practical experiences, *Applied Computer Translation* 1(3), 49-59.

[3] M. King and K. Falkedal (1990), Using Test Suites in Evaluation of Machine Translation Systems, *13th International Conference on Computational Linguistics,* Helsinki, 211-216.

[4] Dan Flickering, John Nerbonne, Ivan Sag and Tom Wasow (1987), Toward Evaluation of NLP Systems, *Ms delivered at Session of 25th Annual Meeting of the Association for Computational Linguistics.*

[5] Margaret King (1989), *A Practical Guide to the Evaluation of Machine Translation Systems,* ms, ISSCO, Geneva.

[6] John Lehrberger and Laurent Bourbeau (1987), *Machine Translation: linguistic characteristics of MT systems and general methodology of evaluation,* Amsterdam, John Benjamin.

[7] John R. Pierce and John B. Carroll (1966), *Language and Machines – Computers in Translation and Linguistics (Alpac Report),* Washington DC.

[8] John A. Hawkins (1985), *A Comparative Typology of English and German: unifying the contrasts,* Croom Helm, London.

[9] R. Lee Humphreys (1990), User-Oriented Evaluation of MT Systems, *DLL Working Papers in Language Processing* 16, University of Essex.

[10] Peter Pym (1990), Simplified English and Machine Translation, *Papers from Tecdoc '90,* Consert, London.

[II] G. van Slype (1982), Conception d'une méthodologie générale d'évaluation de la traduction automatique, *Multilingua* 1(4), 221-237.

[12] Muriel Vasconcellos (1989), Long-term Data for an MT Policy, *Literary and Linguistic Computing* 4(3), OUP, 203-213.

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

AUTHOR

R.Lee Humphreys, Essex MT/CL Group, Department of Language and Linguistics, University of Essex, Colchester, Essex CO4 3SQ, UK