[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland;
ed. Kirsten Falkedal (Geneva: ISSCO).]

User-Oriented MT Evaluation and Text Typology

R. Lee Humphreys
Department of Language and Linguistics
University of Essex

Abstract

A brief survey of user-oriented MT evaluation methodologies suggests that they all suffer drawbacks in terms of *time* and/or *generality* and/or *user interpretability*. The time and cost of evaluation, coupled with system prices, make it increasingly less likely that full-scale pre-purchase evaluation by a single potential user will be economic. It is clearly desirable that MT systems should be "generically benchmarked" in some way to agreed standards. Regardless of the particular evaluation metrics used, generic benchmarking requires corpora of typed texts as test material, presupposing a shared notion of text type. Document structure, which is increasingly subject to international standardisation, provides a useful basis for text typing in this context. Furthermore, elements of a given document structure can be associated with linguistic diagnostics usable in statistical measures of prototypicality.

1 User-Oriented Evaluation

Potential users of an MT system want to know whether it is likely to be an economical factor in the translation of their texts.

Test methodologies that we know about use either prototypical samples of end-user texts or artificial texts constructed by designers and developers, such as a suite of test sentences designed to exhibit linguistic constructions on a "one-per-sent" basis (Flickinger, 1987).

From an end-user perspective, artificial texts are not necessarily the ideal basis for system evaluation:

- For a linguistically naive user, specifying that System X translates gapping constructions successfully is not informative. Even if s/he is shown an example of a gapping construction, it is not immediately obvious what strings are to count as gapping constructions.

- Even if the user knows what a gapping construction is, s/he typically won't know its frequency in the type of natural texts s/he is interested in. It is not easy to find this information, since that presupposes that we have parsers capable of finding instances of all constructions of interest. If we did, we wouldn't be running test suites on them in the first place.

- Any artificial text is likely to be incomplete:

  — Through oversight or linguistic ignorance, it may fail to address a particular construction.

  — Although artificial texts generally attempt to assess performance on combined constructions (e.g. Dative + Passive), it is unlikely that all combinatory possibilities will be explored.

  — The construction of artificial texts for *monolingual* NLP systems is not easy; the construction of such texts for *multilingual* systems is likely to be even more difficult. Examination of MT literature suggests there is only a fragmentary typology of translation problems available.

Unfortunately, there are also problems with evaluation procedures that use natural texts. Two sorts of procedure have been investigated in the literature:

Declarative Evaluation where an attempt is made to measure the absolute performance of a system by scoring its output with respect to various quality dimensions (such as Accuracy, Intelligibility, Style etc) (Pierce & Carroll, 1966). The potential user may have difficulty in determining what the significance of a given result in this type of evaluation might be.

Operational Evaluation where an attempt is made to measure the cost-effectiveness of a system as a factor in the overall translation process (King & Falkedal, 1990; King, 1989; Humphreys, 1990; Lehrberger & Bourbeau, 1987; van Slype, 1982; Vasconcellos, 1989).

It is not the purpose of the present paper to determine which of the two natural text evaluation methodologies is to be preferred nor, indeed, whether either is to be preferred to artificial text methods. What is clear is that they are extremely expensive to conduct. For example, a rather rudimentary declarative evaluation of a PC-based MT system conducted recently at Essex cost considerably more than the hardware+software package itself. Detailed consideration of operational evaluation suggests that reliable results would require a considerable amount of materials preparation, user training, data collection and data analysis (Essex MT Evaluation Group, 1991). The idea that prospective individual users would find it cost-effective to carry out their own detailed evaluation really belongs to the heroic age of MT rather than present realities.

What clearly is important is some methodology for identifying text types in evaluation. For a single individual user, it is enough to select "representative" samples of some archive of source texts that had to be translated in the past, the assumption being that much the same mix of text types will have to be translated in the future. But such a corpus is only extensionally characterised; an intensional characterisation of sample corpora is needed if other users are to have any hope of relating results to their own text translation requirements.

The availability of a text typology is also relevant to the interpretation of results obtained on artificial texts: their significance depends on frequency information, and frequency must be relativised to text type.

2 Text Typologies

Like all entities, texts have intrinsic and extrinsic characteristics. The latter are largely matters of origin and function, whereas intrinsic characteristics encompass syntax, morphology, lexis and so on – i.e. content.

Extrinsic characterisations have received considerable attention in the literature. For example, de Beaugrande (1980) categorises texts into (not necessarily exclusive) functional classes: descriptive, narrative, argumentative, literary, poetic, scientific, didactic and conversational. The categories are clearly related to those traditionally adduced in rhetoric: narration, description, exposition and argumentation. More recently, Dillon and McKnight (1990), using a repertory grid approach[1], obtained evidence that their subjects construe texts in terms of three broad attributes: *why* read them? (e.g. for professional or personal reasons), *what* type of information do they contain (e.g. technical or non-technical), and *how* are they read? (eg. serially or non-serially). Clearly the *how* and *why* attributes are both text extrinsic.

MT and NLP literature has tended to focus on intrinsic text characteristics. Arnold (1988) notes 5 "inherent properties" of texts which can be subject to constraint in any given text type:

Semantic Domain (domain of discourse/subject field) .

[1] Repertory grid analysis was developed as a means of eliciting information on how subjects construe their world. Given a set of elements - in this case various texts - subjects are invited to generate a construct (a dimension of comparison) which would meaningfully discriminate between these elements e.g. Serious vs Trivial. The technique provides a means of quantifying subjective classification schemes.

Overall Discourse Type “Texts may be restricted to those that have a particular internal format or structure, for example, business letters, newspaper stories...”

Discourse Structure “Texts may be such that they exclude, for example, pronominal references outside the sentence or paragraph; that headings may always be noun phrases; that ‘descriptive’ and ‘imperative’ sections of text may be clearly separated.”

Syntax and Morphology “The range of syntactic constructions may be limited such that they exclude all but declarative and imperative sentences; restrict NN compounds to those which are listed as single items; limit the kinds of co-ordination that are allowed.”

Lexis “The vocabulary used may be limited in terms of the number of distinct words that can be used, or in terms of the range of uses of each word....”

As Arnold points out, if the language of texts thereby restricted is used systematically and productively by some language community, then one has a *sublanguage* (Kittredge, 1987) – an attractive domain for MT applications.

Arnold's second inherent property – overall discourse type – draws attention to the fact that many classes of text have a fixed and recognisable internal structure, also known as *logical structure.* At its lower level, such structure may simply be format information, corresponding to traditional text divisions e.g. the division of books into chapters, with bibliographies and index at the end. However, text-structural categories may also be highly intensional e.g. a text-book may contain elements which are considered to be examples, and others which are considered to be exercises. A given class of text may be identified by its internal structure: a text-book is a text entity which contains amongst its text elements both examples and exercises.

Recent years have seen the growth of interest in descriptive text-markup which seeks to identify the logical structure of text prior to its further processing. In particular, the recent ISO 8879 standards characterises a human-readable markup language – SGML – intended for use throughout the document processing industry. It has two main components: the provision of a syntax for describing the structure of a class of documents – the Document Type Definition, and the provision of a related syntax for marking up the actual structure of a document instance of some such class or type. (For detailed presentations of SGML see Bryan, 1988; Goldfarb, 1990)

Formally, the DTD appears to be a regular grammar on document elements (e.g. “chapter”), having as its most primitive elements sequences of character data (e.g. the element “sentence” might be identified with some arbitrary length string of characters). Various syntactic devices, such as string variables and end-tag omission, enable the document writer to minimise markup.

The function of a DTD is to ensure that any given document has a structure conformant to its intended type e.g. to ensure that a sales brochure is made from a sequence of elements compatible with the the sales brochure DTD. It is assumed that marked-up documents will be parsed relative to a DTD (a "document grammar") using an SGML parser software tool; such a parser might also be linked directly to a document-formatting package.

DTDs are complex entities; it is assumed that they will normally be constructed by specialist document designers. Although DTDs themselves are not ISO-standardised, it has always been envisaged that many DTDs will be treated as public or semi-public items. For example, it was envisaged (and it is rapidly becoming the case) that publisher's associations etc. would make available DTDs for documents within their domain of interest, such as text books. ISO has provided guidelines for identification of public DTDs etc. which are expected to be registered with appropriate authorities.

This ongoing process of SGML standardisation in document preparation is likely to result in an increasing tendency towards a shared (e.g. European) notion of what the structure of various conformant business documents must be; we can expect to see widely shared notions of what constitutes the internal structure of a user manual or a sales brochure. In such circumstances the MT community might benefit by selecting various publicly defined document types and then using samples of such types (text instances) for generic MT benchmarking. The selected document types must, of course, be reasonable candidates for MT.

The underlying assumptions of this strategy are:

- The text types identified are reasonably natural - they correspond to types routinely recognised in (say) the business world
- The internal structure of each text type is non-negligible and sufficiently rich for its individuation.
- Instances of a given text type in a given subject field share more linguistic similarities than instances of different text types in the same subject field.

To give an idea of the stability of internal structure in existing documents, we conducted a simple straw poll. We asked three sources – 2 Colchester-based engineering manufacturers and a University technical department – to provide us with some samples of "product information brochures"[2]. We assumed that this was a natural text category and hypothesised that it

[2] All products were to be manufactured goods for the professional rather than the consumer sector.

would necessarily contain the element “Product Description” and, optionally, “Product Specification” (a simple attribute-value structure), “Company Information”, “Ordering Information” and so on. We received about 21 brochures, but of these 5 were not directly about particular products or ranges of products and were discarded[3]. Of the remaining 17, 4 came from the technical department (describing recording equipment, waveform analysers and laboratory furniture), 4 from a bearing manufacturer (all describing various ranges of machine bearing) and 9 from a machine-tool manufacturer (all describing various centre lathes or control units). 15 of these brochures were clearly conformant to our expectations, containing sections that could clearly and unproblematically be described as product descriptions. The following is a typical example of material found in such an element:

> The unit is constructed from welded tubular square section steel and fitted with adjustable feet. The finish is tough epoxy coating, standard colour black. The vacuum surface is melamine faced to give a flat, rigid easily cleaned surface. All sliding surfaces are chrome plated or stainless steel.

One bearing brochure did not really contain a product description at all; the text gave a general account of bearing specifications and bearing standards, relegating all detailed product information to a specification table. In the remaining brochure, information about a particular bearing product was embedded – albeit as a single block – in a larger text describing the failings of other bearings as produced by other manufacturers.

To summarise this mini-survey: product information brochures really do contain a product description text element.

3 Diagnostics

We have suggested that functionally identified text types such as product information brochures have functionally identifiable sub-elements. It is a reasonable assumption that these sub-elements have some particular linguistic content which mediates their particular communicative function. A Product Description, for example, is likely to be recognisable as such in virtue of two factors:

1. An underlying text-linguistic structure which typically consists of a series of taxonomic (“The PRODUCT is X”), constituency (“The PRODUCT is X”) or property statements (“The PRODUCT has Y”):

[3] For example, one was a booklet entitled “Flexible Manufacturing at Colchester Lathe Company” and another was clearly a reprint of a trade advertisement

> The machine is of modular construction with a massive bed mounted on a substantial base. The 45 degree inclined carriage is guided along hardened steel sheers.... The main bearing has a guaranteed roundness accuracy better than 2 microns.

2. Characteristic lexico-syntactic realisations of such statements.

Biber (1988, 1989) has developed a five-dimensional model of text, where each dimension corresponds to lexical and syntactic features that co-occur in texts according to their particular communicative function. For example his Dimension 2 distinguishes “Narrative versus Non-Narrative concerns”; narrative texts are said to exhibit the following features:

past-tense verbs
3rd person pronouns
perfect-aspect verbs
public verbs (e.g. speech-act verbs)
synthetic negation
present-participial clauses

whereas non-narrative texts will tend to exhibit present-tense verbs and attributive adjectives. The linguistic features used in constructing the dimensions are drawn from 16 grammatical categories e.g. tense/aspect markers, place/time adverbials, passivisation, specialised verb classes, coordination etc. These features were machine-counted on the LOB and London-Lund corpora (representing some 23 different genres, including spoken texts) and then associated by means of factor analysis.

In his most recent work (Biber, 1989) Biber has identified 8 *functional* types of text, where each type represents a grouping of texts which are similar in respect of their dimensional characterisations as determined by cluster analysis. To these types – which seem to cut across text genres – Biber has assigned such labels as “Situated Reportage”, “Learned Exposition”, “General Narrative Exposition” etc.

Biber’s work is somewhat unusual in that, having first identified a number of linguistic components of texts, he then constructs co-occurrence groupings of linguistic functions related to communicative functions and finally a functional text typology from those groupings. By contrast, the usual strategy is to start from function and then work back towards the linguistic substrate of function.

Biber’s inferred text types are very general, and there is no attempt to deal with the fact that different sub-elements of a given document type – say a user manual – will contain sub-elements such as “Product Description” and “Fault Finding” which have very different communicative functions. For such subelements, we are interested in filling out a highly intensional/functional description (i.e. "Product Description") with a list of linguistic co-occurrence phenomena – the lexico-syntactic realisations of the propositions mentioned above. In fact, it is quite straightforward to come up with a preliminary list (derived by inspection of the cited sample product information brochures):

Present tense
Stative verbs (e.g. have, permit, allow, offer, be, constitute, form, include)
Attributive adjectives
Epistemic modalities (e.g. "it is also possible to describe the tool path on the screen")
Comparative/Superlative adjective forms (e.g. "makes the instrument easier to operate")

We can regard such co-occurrence features as diagnostics i.e. a prototypical "Product Description" text-entity will tend to exhibit high frequencies of these features. The way is open to statistical control of "prototypicality" when selecting text instances of a text element type: select only those instances which exhibit a high frequency of the diagnostic features (and low frequency of complementary features).

3.1 Diagnostics and Frequencies

Diagnostic features "à la Biber" occur with high frequency in a given text type. It was suggested earlier that one reason for questioning the acceptability of artificial texts is the lack of frequency information on various linguistic constructs; have we now admitted that such information really is readily available?

The answer is no:

- Diagnostics are highly partial; they suggest that some constructions occur with high frequency in a given text, and that others occur with low frequency; nothing is said about the frequencies of other constructions.
- Text-typing diagnostics are mono-lingual. They do not encode any expectations about translational phenomena.

In effect, we are looking at ways of linguistically profiling texts rather than any comprehensive statement on the frequencies of all possible constructions of interest.

4 Summary and Conclusions

Users require interpretable generic benchmarking of MT systems. Difficulties in interpreting artificial texts, and ensuring their adequacy as probes of translational prowess, suggest that natural texts have a part to play. Translation performance should be assessed relative to text type – hence sample selection requires a text typing methodology. It is proposed that the notion of document type, entailing an identification of document text sub-elements, provides a good initial basis. Furthermore, intensional descriptions associated with document elements can be associated with simple linguistic diagnostics; these may offer a statistical measure of prototypicality.

References

Douglas Arnold (1990), Text Typology and Machine Translation: An Overview, *Translating and the Computer,* 10, ASLIB, London

Douglas Biber (1988) Spoken and Written Dimensions in English: resolving the contradictory findings, *Language* 62, 384-414

Douglas Biber (1989) A Typology of English Texts, *Linguistics* 27, 3-43.

Martin Bryan (1988), *SGML -- An Author's Guide to the Standard Generalized Markup Language,* Addison-Wesley, Wokingham

De Beaugrande (1980), *Text, Discourse and Process,* Ablex, Norwood NJ

Andrew Dillon & Cliff McKnight (1990), Towards a Classification of Text Types: A Repertory Grid Approach, *Int. J. Man-Machine Studies,* 33, 623-636.

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

Dan Flickinger, John Nerbonne, Ivan Sag & Tom Wasow (1987), Toward Evaluation of NLP Systems, Ms *delivered at Session of 25th Annual Meeting of the Association for Computational Linguistics*

Charles Goldfarb (1990), *SGML Handbook,* OUP, Oxford

Margaret King (1989), *A Practical Guide to the Evaluation of Machine*

Translation Systems, ms, ISSCO, Geneva

Richard I. Kittredge (1987) The significance of sublanguage for automatic translation, *Machine translation: Theoretical and methodological issues,* Sergei Nirenburg (ed.), *Studies in Natural Language Processing,* CUP, Cambridge, 59-67

John Lehrberger & Laurent Bourbeau (1987) *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation,* Amsterdam, John Benjamin

John R. Pierce & John B. Carroll (1966) *Language and Machines – Computers in Translation and Linguistics (Alpac Report),* Washington D.C.

R. Lee Humphreys (1990), User-Oriented Evaluation of MT Systems, DLL Working Papers in Language Processingl6, University of Essex

Essex MT Evaluation Group (1991), Proposal for a Study of Translation Evaluation, DLL Memorandum, University of Essex

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

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