EPOS, XVI (2000), págs. 351-360

PRINCIPLES OF SUBLANGUAGE THEORY AND THEIR RELEVANCE TO MACHINE TRANSLATION

ELENA BÁRCENA & TIM READ
UNED. Madrid

ABSTRACT

This article discusses the different ways of constraining the input to be treated by a Machine Translation (henceforth, MT) system, focusing on the characterization of natural sublanguages according to research on Sublanguage Theory and the way in which their properties benefit the quality of MT output. The problem of designing an MT system for a natural language is seen from the perspective of the efforts necessary to write a formal grammar for that language. This article argues that the benefits of designing and developing a system for the translation of a sublanguage instead of a general language imply a simplification of both the mechanism and the representation to varying degrees. Although gradual improvements in formal grammars will undoubtedly be made in the near future continued heavy reliance on post-editing or Interactive MT is expected.

INTRODUCTION: WAYS TO CONSTRAIN MACHINE TRANSLATION INPUT

This article discusses the different forms of constraining the input to be treated by a Machine Translation (henceforth, MT) system, focusing on the characterization of natural sublanguages according to research on Sublanguage

Theory and the way in which those properties improve the quality of MT output.

The question of input restriction emerged as a consequence of the need to reduce the large and variegated linguistic knowledge that must be inserted into a system, and also to simplify the process of translation. Nowadays, given state-of-the-art MT, in order to capture and utilise lexical, grammatical, semantic, contextual, and real-world knowledge, it is apparent that the language and its referential world, as well as the elements that compose the corresponding communicative scenario of the system must be somehow limited in comparison to general languages. There is general agreement that in order to obtain high-quality output, a system that is not limited in its coverage must necessarily be highly interactive, by the same token that only when a system limits its coverage can it be fully-automatic (Lehrberger & Bourbeau, 1988, p.128).

The most widespread practice of input restriction throughout the history of MT has been the general-purpose corpus-based approach, which limits the object of the translation to a written sub-field both in experimental and operational systems (e.g., GAT [physics texts], CETA [mathematics and physics texts], ALP [Mormon ecclesiastical texts], METAL [technical documents], SPANAM & ENGSPAN [medical and public health documents], Eurotra's prototype [information technology] [Slocum, 1984, pp.6-9]). However, «their [...] limitations are regarded as temporary: extension to other subject areas is anticipated» (Hutchins & Somers, 1992, p.323), which means that the system's design and performance do not gain substantial advantage from the constraints.

The other tactics used to constrain translation input have been basically dictated by the system, for example, the use of restricted syntax and controlled vocabulary (Hutchins & Somers, 1992, p.151-152), which consists of the imposition of structural and lexical constraints on the authors of the texts which are going to be translated, i.e., texts can only be expressed adhering to a pre-defined set of constructions and a limited set of words. However popular these approaches may be (e.g., the use of Systran by Rank Xerox [Hutchins & Somers, 1992, p.152] and TITUS [Hutchins, 1986, pp.293-294] respectively), linguistic manipulations of the input limit the scenarios in which the system can be used since, for instance, trained authors are required for their composition, and the naturalness of the language is inevitably affected (to such an extent that authors talk about, for instance, «Systran French» [J.C. Sager, 1986, p.166]). Furthermore, they are not the ideal solution to the problems of whole language formal description (Laurian, 1984, p.238), at least in basic research. The ultimate aim of translation is to transfer into a given language the

words and/or communicative intents from a text (Newmark, 1981, p.62-63), and not to intervene in the author's task or distort them.

There is a common-sense convenience entailed in attempting MT from the perspective of real-world demands for translation. Indeed, the reason for attempting the application of computers to the translation of texts is to attend to the urgent needs of political and military bodies, scientists, technologists, businessmen, and professionals in general to read a growing number of documents and/or communicate constantly in languages they do not know. The hope is that, in the long term, computers will relieve humans of the task of translating high-volume (and frequently routine) documentation, so that they can concentrate on more creative and interesting translation tasks (Tucker, 1987, p.28).

There is an increasing awareness that «the commercial success of machine translation in the foreseeable future likely depends on the possibility of writing [...] grammars for texts in particular fields» (Kittredge & Lehrberger, 1982, p.3). Since the language in which these texts are expressed can often be defined as a *sublanguage*, the obvious connection between research in the fields of MT and of Sublanguage Theory was finally made, and resulted in practice in the creation of the first operational fully-automatic high-quality translation system: TAUM-METEO (Chandioux & Guérard, 1981).

There is an important distincion to be made between sublanguage-based MT systems, which was first observed by Lambropoulou (1989, p.55) (who, incidentally, prefers the term *restricted language*):

— MT systems with natural sublanguages: These sublanguages imply a series of restrictions and deviations from their respective whole languages which have been developed exclusively by the users of these sublanguages according to their expressive needs. No changes or manipulations are imposed on them at all during the construction of an MT system intended to translate texts written in these sublanguages. Quite the contrary, all grammars and lexica of an MT system of a natural sublanguage exist before they are even considered to be applied for that system, and have no relation with its computational needs.
— MT systems with pre-edited sublanguages: Constraints may emerge gradually from the act of communication without any effort on the part of humans («natural sublanguages» or «naturally restricted languages») or rules may be invented by a committee to facilitate written communication between specialists. These sublanguages — usually called «constructed restricted languages» or «pre-edited

sublanguages» [1]— are the result of imposing further constraints on naturally restricted languages. Restrictions can be established by human users alone or introduced interactively after being required by the machine.

The most positive effects which have been obtained through artificial restriction are better readability of the input text and a considerable improvement in the quality of the MT output so that post-editing is often unnecessary. As for the disadvantages of these pre-edited sublanguages, they are the same as for the lexicallyand structurally constrained languages introduced above, such as the artificiality of style. Some authors have criticised language pre-edited by computer as somewhat stiff or artificial, although research for «naturalizing» it is in progress (Sager, 1990, p.1). Another factor to take into account is the cost-effectiveness of the constraining process. Therefore, Cyre's (1985, p.128) advice for those building these systems is that they must be «readable without training and writable with little training». In order to apply this strategy to a system the positive aspects of style (clarity and consistency) and their computing consequences must be weighed up against these disadvantages.

Principles of Sublanguage Theory

One of the first relevant statements in the field of Sublanguage Theory was Harris' (1968, p.152) view of a whole natural language as a compound of subsystems (in the mathematical sense): «certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it». This novel approach distinguished the existence of sublanguages within a whole natural language on the basis of their closure under transformations such as negation, question formation, or clefting. The type of closure property proposed by Harris is a necessary condition, but not sufficient, for identifying a sublanguage: «if a sublanguage can be any subset of sentences which is closed under the transformational operations, this definition could identify a very large number of linguistic subsets as sublanguages» (Kittredge, 1982, p.110).

Although Harris did not mention the role of subject-matter in the delimitation of sublanguages at the time, it was later observed that specialists in a particular technological, scientific, or professional sub-field tend to share lexical, semantic, syntactic, and discourse usages. Consequently, a new significant definition of sublanguage emerged: «the language used by a

[1] The term «artificial (sub)language» is not to be used for this concept to avoid ambiguity since it specifically refers to a different type of «man made» language.

particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialised occupation» (Bross et al., 1972, p.1303). This phenomenon is due to the nature of human languages as cybernetic communicative tools, i.e., to both their capacity and tendency to adapt themselves to the different communicative situations in order to maintain maximum efficiency in all their uses.

A community of speakers is normally linked by some common knowledge about the domain which goes beyond that of speakers of the standard language. This opinion is supported by the circumstances in which sublanguages generally emerge: scientists, technicians and people in general establishing communication about a specific subject matter in a professional or erudite way gradually begin to manipulate and adapt the rules of the language they use in accordance with their communicative needs.

The more specialised and structured the content of the domain is, the more homogeneous the language, i.e., the more rigid the barriers are between what can and cannot be said. This is why it is easier to recognise a technical or scientific sublanguage as compared to, for instance, a journalistic sublanguage and, in fact, the sublanguage phenomenon is more frequently observed in technical and scientific communication. Nevertheless, as communicative needs continue to modify natural languages, existing sublanguage boundaries will become more and more blurred.

However, it must be admitted that for some sublanguages the relevant community of speakers is not well defined, especially for written sublanguages where access to the texts is relatively free. According to many critics, this fact does not diminish in the least their entity as sublanguages. For these sublanguages (e.g., newspaper job advertisements) the authors agree with Kittredge (1982, p.111) that «different subtypes evolve which are oriented to different categories of user, dependent on their level of expertise. [...]. There is every reason to call them different sublanguages for purposes of computational treatment» provided that they make distinctive and homogeneous uses of linguistic forms and phenomena.

A further step in the characterisation of sublanguages was the identification of notable restrictions in comparison with standard languages, hence Kittredge & Lehrberger's (1982, p.2) definition of *sublanguage:* «those sets of sentences whose lexical and grammatical restrictions reflect the restricted sets of objects and relations found in a given domain of discourse» (*our emphasis*).

For example, the word *boca* in odontology literature only means «mouth», whereas in standard Spanish it has also other senses (e.g., «entrance», «side», etc.); and neither question tags nor certain verbal tenses

appear at all in, say, English newspaper job advertisements. Those standard language meanings or the rules relative to the construction of the entities above would not be required by the lexicon and grammar of the corresponding sublanguages.

For these reasons, many authors considered a sublanguage to be a subsystem within a given language partly because it is common knowledge that any natural language is a composite of subsystems, which are the result of communication between specialists. Sublanguages comprise lexemes, morphemes, and constructions that are easily identifiable with a specific human language. One can immediately recognise, for instance, if a cooking recipe is written in English or Spanish. For them, the major feature about sublanguages is that they are limited in reference to a specific subject domain, and accordingly use fewer words and constructions than the corresponding general language. As Lehrberger (1986, p.20) comments, it is often believed that sublanguage grammars would derive from general languages simply by deleting a number of rules that are not relevant.

In later studies, the notion of sublanguage has also been applied to the language of certain technical areas, whose sentential units may be shortened to such an extent for rapid communication and maximum precision that they cannot be considered as grammatical utterances in the standard language; for example if they have determiners, auxiliaries or main verbs omitted: both deletion of general rules, modification of others to cover particular cases, and insertion of new ones are usually required for a complete description of a sublanguage.

The application of a standard grammar rule to a particular sublanguage can result in an ungrammatical sentence in that sublanguage and vice versa (e.g., direct object omission in transitive verbs is unacceptable in standard French but common in French cooking recipes; in Spanish job ads any part of speech can be capitalised in order to stand out visually). Therefore, in general, the relationship between a sublanguage and its corresponding standard language is not best described as one of inclusion (or complete independence, but as one of semi-autonomy or intersection (Lehrberger, 1986, p.23); in N. Sager's words (1986, p.4): a sublanguage grammar is not necessarily «a subset of the grammar of the parent language and in fact intersects it».

This is an important point to bear in mind while extracting the grammar of a sublanguage for the purpose of an MT system: although the similarities between a sublanguage and its whole language as pointed out above (morphological rules, part of the lexicon, syntactic constructions, etc.) must not be forgotten, some rules of a given general language are not to be found in its sublanguages (e.g., colloquialisms), while the application of other rules of

certain sublanguages would result in ungrammatical sentences in the general language (e.g., the omission of the article), or at least would be stylistically undesirable (e.g., a mixture of numbers and letters).

The obvious consequence of sublanguage deviations from and restrictions on standard language is that a lexicon and grammar of the latter would not provide a description of all, and only, the content of the former, hence the absolute necessity for sublanguage analysis when wanting to study the language of certain specialised texts and replacement or additional rules.

For example, *overhead* is an adverb in standard English and a preposition in the corresponding sublanguage used by airline pilots, as in *Our route tonight takes us overhead Paris*; and in the medical diagnosis sublanguage an additional transitivity pattern is needed for the verb *to present* as in *The patient presented with the following symptoms.*

It must be stressed that what makes defining the particular lexicon and grammar of a sublanguage an achievable task is not only its homogeneity, the reduction in lexical variety and richness, the restricted use of certain categories and constructions, the divergence from standard language, the use of special knowledge, or standardised terminology, but the fact that units and phenomena at all linguistic levels generally follow patterns of usage, such as strict terminological selection to denote entities or relations in the referential sub-world, and word co-occurrence patterns. Kittredge (1982, p.110), for instance, comments that the semantic limitation of the domain of the discourse is not a sufficient condition for the identification of a sublanguage: «what is required [...] is that there be shared habits of word usage on the part of the speakers». Similarly, Sager (1982, p.9) points out that: «the research papers in a given science subfield display such regularities of co-occurrence over and above those of the language as a whole that it is possible to write a grammar of the language used in the subfield».

This property of systematicity in all its manifestations is one of the natural consequences of the classificatory role of language, which causes sublanguages to reflect the strict organisation of the part of the real world which they describe (while «the whole language imposes only the broadest structuring upon our perception of the world» [Harris, 1982, p.235]).

Finally, it is interesting to point out that sublanguages share most of the universal properties with standard languages, such as unlimited generative capacity (Moskovich, 1982, p.192) and completeness, although sublanguages have only relative completeness because it is a limited subfield of reality which each one describes. The capacity of a sublanguage to «describe by its means any imaginable situation, any message in the area of reality which it serves as a language» (Moskovich, 1982, p.193) is the property which often determines the utility of its practical real world application.

CONCLUSIÓN

The problem of designing an MT system for a natural language can be clearly seen from the perspective of the efforts necessary to write a formal grammar for that language, and it must be said that the benefits of designing and developing a system for the translation of a sublanguage instead of a general language imply a simplification of both the mechanism and the representation to varying degrees.

After commenting on the properties of sublanguages in the previous section it is evident that one solution lies in restricting one's attention to these peculiar vehicles of verbal communication. In fact, there is practically total agreement among critics (Hutchins, 1986) that at the present stage of development of Computational Linguistics, work on the construction of practical systems for automatic translation should concentrate on sublanguages, as the best practical results are expected to be achieved from texts which are naturally constrained both in form and content.

The characterisation of sublanguages as natural languages is currently questioned by authors like Lehrberger (1986, pp.35-36) and Marsh (1986, pp.114-116), who have observed how the borders between «natural languages» and «artificial (programming) languages» are becoming more and more blurred. However, sublanguages are to be included here among the former on the basis that «natural languages» have been originated by social convention or standardization committees (to facilitate communication efficiency), while the features of «query languages», «programming languages» and other linguistic communicative means are dictated by the machine.

MT observers continuously offer defences of Sublanguage-Based MT. As Lehrberger (1982, p.82) says, «it is within the domain of sublanguages that automatic translation appears to be practicable». While analysing linguistic problems for MT, Hutchins (1986, p.325) also concludes from a practical view that, in order to overcome them, «much can be achieved with restrictions of systems to particular sublanguages», rather than aiming at more linguistically sophisticated systems. Similarly, Boitet (1990, p.30) says that the status of sublanguage is desirable for the object of translation even if post-editing of raw output is allowed. By the same token, Lambropoulou (1989, pp.103-104) states that those languages that lend themselves to MT are «those used in instruction manuals, abstracts and texts taken from specialised literature and intended for a specific purpose clearly determined in them, usually communication of information». Otherwise, although gradual improvements in formal grammars will undoubtedly be made, in the near future continued heavy reliance on post-editing or Interactive MT is expected.

BIBLIOGRAPHY

BOITET, C. (1990): Towards Personal MT: general design, dialogue structure, potential role of speech. In *COLING-90* (**3**). Yliopistopaino (Helsinki) pp. 30-35.

BROSS, I.D., P.A. SHAPIRO and B.B. ANDERSON (1972): How information is carried in scientific sub-languages. In *Science* **176**. pp. 1303-1307.

CHANDIOUX, J. & M.F. GUERARD (1981): Mètèo: un système. L'èpreuve du temps. *Meta* **26**. pp.18-22.

CYRE, W. (1985): The design of a restricted sublanguage. In *Theoretical Approaches to Natural Language Understanding: a Workshop at Halifax* (Nova Scotia). pp. 128-130.

GRISHMAN, R. & R. KITTREDGE (eds.) (1986): *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Hillsdale. Lawrence Erlbaum Associates.

HARRIS, Z. (1968): *Mathematical Structures of Language*. New York. Wiley (Interscience).

— (1982): Discourse and Sublanguage. In Kittredge & Lehrberger (eds.). pp. 231-236.

HUTCHINS, W.J. (1986): *Machine Translation: Past, Present, Future*. Chichester. Ellis Horwood.

HUTCHINS, W.J. & H.L. SOMERS (1992): *An Introduction to Machine Translation*. Cambridge. Academic Press.

KITTEREDGE, R. (1982): Variation and Homogeneity of Sublanguages. In Kittredge & Lehrberger (eds.). pp. 107-137.

KITTREDGE, R. and J. LEHRBERGER (eds.) (1982): *Sublanguage: Studies of language in restricted semantic domains*. Berlin. De Gruyter.

LAMBROPOULOU, P. (1989): *The Application of Restricted Languages in Machine Translation*. M.Sc. Dissertation. UMIST. Manchester.

LAURIAN, A.M. (1984): Machine Translation: What Type of Post-Editing on What Type of Documents for What Type of Users. In *COLING-84*. Stanford (Ca.). pp. 236-238.

LEHRBERGER, J. (1986): Sublanguage Analysis. In Grishman and Kittredge (eds.). pp. 18-38.

— (1982): Automatic Translation and the Concept of Sublanguage. In Kittredge & Lehrberger (eds.). pp.81-106.

LEHRBERGER, J. & L. BOURBEAU. (1988): *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation. Linguisticae Investigationes: Supplementa* **15**. Amsterdam. John Benjamins.

MARSH, E. (1986): General Semantic Patterns in Different Sublanguages. In Grishman & Kittredge (eds.). pp. 103-127.

MOSKOVICH, W. (1982): What is a sublanguage? The notion of sublanguage in modern Soviet linguistics. In Kittredge & Lehrberger (eds.). pp. 191-205.

NEWMARK, P. (1981): *Approaches to Translation*. Oxford. Pergamon Press.

SAGER, J.C. (1986): Conclusions. In *World Systran Conference*. pp. 161-166.

SAGER, N. (1986): Sublanguage: Linguistic Phenomenon, Computational Tool. In Grishman & Kittredge (eds.). pp. 1-17.

— (1982) Syntactic Formatting of Science Information. In Kittredge & Lehrberger (eds.). pp. 9-26.

SLOCUM, J. (1984): Machine Translation: its History, Current Status, and Future Prospects. Working Paper LRC-84-3. Linguistics Research Center. University of Texas (Austin). pp. 546-561.

TUCKER, A.B. (1987): Current strategies in machine translation research and development. In S. Nirenburg (ed.). *Machine translation. Theoretical and methodological issues*. Cambridge University Press. pp.22-41.