Usage anomalies and internationalization

Alfs T. Berztiss
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
E-mail: alpha@cs.pitt.edu
and
SYSLAB, University of Stockholm, Sweden

Abstract

Communication can fully succeed only if the recipient of a message gives the same interpretation to the message that its sender intended. The achievement of successful communication becomes problematic in an international setting in which the senders and receivers of messages have different linguistic and cultural backgrounds. The different linguistic backgrounds manifest themselves as spelling and syntax anomalies. Cultural differences result in a context switch, which can substantially change the meaning of a message. We suggest that messages be explicit and robust. Explicitness means that there is little dependence on a cultural context; robustness means that potential sources of linguistic misinterpretation, such as double negatives, are avoided. We point out research directions that aim at dealing with the problems associated with internationalization.

1 Introduction

It is largely a linguistic concern to ensure that the intended meaning of a message is transmitted to the recipient of the message. Very often there is not even an awareness that we are misinterpreting a message. Now that English is becoming the international language of business and science, misinterpretation of messages becomes a serious concern. The aim of this paper is to examine the communication problems that arise with the internationalization of language, and to suggest the research needed to deal with these problems. Some of the problems arise even in communication between native speakers of a language, primarily due to increasing diversity in the cultural backgrounds of the people between whom messages pass. Our discussion will relate primarily to business information systems.

The problems raised by adaptation of software for different markets have been discussed quite extensively [1–6]. Some relate to usage differences in different languages (English and Japanese), others to cultural differences between countries that share the same language (England and Canada, France and Canada, Portugal and Brazil). An ingenious system of communication between peoples that speak different languages are Chinese ideograms. We are now reaching a stage at which English words are becoming ideograms that are being put together by non-native speakers of English for expressing their communications. The results do not always obey the norms of English usage, but in time they can change existing usage patterns.

Two questions arise. First, if the meaning of a message is not compromised by non-standard usage, does usage matter? An analogous situation existed in the Middle Ages with Latin. It was then the universal language for many domains, including science and religion, but it was not the "standard" Latin of Caesar or Cicero. Second, how can we prevent non-standard usage, or, when non-standard usage has altered the intended meaning of a message, how are we to restore the intended meaning? Here we shall consider the second question alone. Our concern, then, will be to find a way in which people with different cultural and linguistic backgrounds can use English as a means of communication without their messages losing too much of their intended meaning.

In Section 2 we shall discuss three types of anomaly — of spelling, of syntax or of sentence construction, and of context switch. By context switch we refer to the change of meaning a message undergoes when it is interpreted under two different frames of reference. Sections 3–5 will suggest ways of dealing with these anomalies. We do not offer ready solutions; we merely point out directions for in-depth research. Section 6 is a summary of our findings, which includes a research agenda for the future.

Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)
0-7695-0680-1/00 $10.00 © 2000 IEEE

2 Types of linguistic problems

Even native speakers of English make spelling errors. Two common ones are the confusion of *principal* and *principle*, and the use of *ie* instead of *ei* in some past tense constructions, such as writing *recieved* instead of *received*. Now, *recieved* presents no problems because there is no such word, but if a computer program is to determine whether *principal* means what it says or is a substitution for *principle*, then a sizable region of a semantic context space has to be examined. What matters are just those deviations that make it difficult for the non-native speaker to understand a word, and may prevent looking the word up in a dictionary, i.e., cases in which a spelling mistake has resulted in a new legitimate word.

Languages differ greatly in their syntactic rules. Thus, the means used in some languages to deal with number and tense are quite different from those of English. There can also be an absence of articles, or an article can be an integral part of a noun, or articles are used in ways different from those in English. Particularly troublesome are double negatives. In some languages they strengthen the negative, in others they neutralize it. It is frustrating not to be able to tell whether something must be done, or must at all costs be avoided.

Most comm unications with which information systems have to deal follow what we shall call a wh-pattern. This means that in addition to defining *what* is their target of concern, they are also interested in locating it in time (*when*) and space (*where*), and in identifying participants (*who*). Let us look at an example: "The estimated sales figure for Stockholm for 2001, as supplied by the head office, is SEK 8,000,000." Here the what-component is "The estimates sales figure is SEK 8,000,000," the where-component is "Stockholm," and the when-component is "2001." The who-component is "the head office," and here it identifies the source of the information, but it could also identify the user of information, as in "The regional office is basing its budget for 2001 on a sales estimate of SEK 8,000,000 for Stockholm." In some languages the order of the wh-components is strictly fixed. In others, such as English, there are no ordering rules, and the order may be used to indicate the relative significance of the wh-components. Thus, the who-component is emphasized in "The head office estimates sales for Stockholm for 2001 to be SEK 8,000,000."

This brings us to context switch. When somebody with a good sense of English carefully arranges the wh-components of a communication so that particular emphasis is put on location, say, and the message is interpreted by somebody for whom the wh-components always follow the same strict order, the intended emphasis is likely to be lost. Indeed, it may be lost even on a native English speaker with no ear for subtle stylistic hints. This suggests that business language be explicit and robust. Explicitness requires that dependence for interpretation of a message on a particular cultural context be minim al. Robustness implies avoidance of double negatives and other potential sources of confusion. We have given a few hints on how to go about this [7, p. 136], e.g., we recommend that synonyms be avoided because the denotation of a single concept by two terms leaves a reader uncertain whether the two terms do in fact relate to just one concept. Robustness with regard to cultural contexts requires that metaphors be avoided. For example, red stands for danger in some cultures, but for celebration in others [8]. However, as much as we may wish the explicitness and robustness recommendation to be followed, reality will be different, and we have to find ways of dealing with this reality.

3 Spelling control

Spelling checkers can remove a lot of uncertainty, but experience with papers I have had to referee suggests that they are not always used. Moreover, a spelling checker merely lists words that are not in its dictionary, which can make it difficult, particularly for a non-native speaker, to make a correction.

The most common spelling mistak es are letter substitution ("seperate" for "separate"), transposition ("palce" in place of "place"), omission ("personel" in place of "personnel"), insertion ("appartment" for "apartment"), or some combination of the above ("personell" in place of "personnel"). A metric should supply distances between the misspelled word and correct words, and an ideal metric would be based on averages weighted in accordance with the expected frequencies of the types of mistakes listed above. However, the assignment of weights would be very difficult because some errors depend on the native language of the writer. For example, Swedes seem to have an affinity for "appartment". The reader who knows the meaning of "apartment" would have no difficulty in decoding the misspelled word, but the person who has to look up "apartment" in a dictionary faces a difficulty. Because Finnish does not have the letter "b", a substitution of "p" for "b" can easily arise.

In view of the high processing speeds that are now available, our suggestion is that there be no search for a metric, but that a word that is not found in a dictionary be subjected to a battery of transformations. In

Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)
0-7695-0680-1/00 $10.00 © 2000 IEEE

a word with ten letters, only nine transpositions and 250 single-letter substitutions are possible. No system can, of course, be perfect. For example, there may be two omitted letters in the same word. But, if a spelling error has resulted in another legitimate word that does not quite fit in, the transformation system should allow the reader to find the word that was intended. Context may allow ranking of the words when the system generates many candidates. For example, if a Finn writes "pank", then "bank" is more likely to be the intended word than "rank", "sank", "tank", "yank", "pink", etc.

4 Sentence construction

Even a native speaker can easily make a spelling mistake. Native speakers of English have also problems with the order of words. Thus, English is very liberal with regard to word order, but, depending on where "only" is placed in "programmers debug programs," three semantically different statements are obtained. They are: nobody but a programmer does it; this is all that programmers do; nothing but programs get debugged by programmers. Therefore, a system that paraphrases text can point out errors in word order that might otherwise go unnoticed. Such a system would also clarify the meaning of a phrase in which a double negative appears. In case somebody writes "programmers and testers are located in different buildings; they do not communicate," such a system could suggest two interpretations, namely that programmers do not comm unicate with testers, or that the buildings are not directly linked.

Although it seems that the omission or misuse of articles, and the wrong use of number and tense would create frustrating problems, this is not so. When a particular written language puts little emphasis on such matters, and somebody brought up on this language writes "Programmer debug own program," the phrase is easily understood. I have had to deal on a daily basis with material written by Chinese and Americans. Surprisingly and counterintuitively, I have found that although the writing by the Chinese is far from standard English, there is less room for misinterpreting their writing than that produced by Americans. The "Chinese English" is more robust.

This raises the important question of how far the commonly accepted norms of English should be relaxed. Alternatively, should scientific and business writing be governed by their own set of rules. In manufacturing industries, such as the aircraft industry, user manuals and maintenance manuals are being written in a restricted form of English with a very limited syntax and general vocabulary, supplemented with the specialized vocabulary of the domain. Such a language can be used for the communication of facts and of experimental results as well.

However, scientific and business writing is not just for communication of facts or of experimental results. Quite often position papers or business proposals have to be written, where the aim is to persuade the reader to adopt or to support a particular approach. Persuasion needs a richer vocabulary than is offered by some restricted form of English.

We need then a way of assisting a writer not in full command of English syntax to arrive at "standard" prose. Our suggestion is that the writer starts out by creating a conceptual structure and that a text generator coupled to a paraphraser helps the writer to transform this conceptual structure into stylistically acceptable prose. In Section 1 we noted that English words can be regarded as building blocks similar to Chinese ideograms. The conceptual structure would be similar to the structure that results when Chinese ideograms are combined, but its exact form will be determined by the need to couple it to a text generator, and this remains a research topic.

5 Context switch

A dictionary is not always a dependable tool for assisting communication. At an international conference papers presented in French were given simultaneous translation — it took some time to realize that "exploitation systems" were really "operating systems". There is now a realistic appreciation of the difficulty of natural language processing, and we understand that it requires the solving of very many highly specific problems, e.g., problems that may relate to just one particular word. By means of this step-by-step approach we have built up a repertoire of problem-and-solution patterns, which today allow natural language processing of fairly high quality.

Nevertheless, a formidable set of problems remains, and, as one problem is solved, new ones take its place. This is inevitable because the domains in which natural language is applied keep changing. Also, language itself changes, but not everywhere at the same rate. Thus, expressions that have long since disappeared from English as used in Britain, still persist in the English of India. Hence outdated usage patterns cannot be phased out of language processing systems.

The ideal solution for avoiding misinterpretation due to context switch is to insist on explicitness and robustness, but, particularly in prose aimed at persuasion, metaphors will be used, and they can be misinterpreted. One solution is to codify the different cultural frames of reference, and to substitute metaphor

Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)
0-7695-0680-1/00 $10.00 © 2000 IEEE

for metaphor. Such codification would be an immense project unlikely to become practicable in the near future.

An alternative that may be somewhat easier to implement is to flag expressions that do not meet the explicitness and robustness requirements. Recognition of double negatives is quite easy. So is the recognition of the metaphoric use of "red" in the phrase "uncommented prose makes a tester see red," but not so in "heavy commenting of a piece of code is a red flag for a tester" — "to see" a color is not part of expected usage, but "red" can be applied as an adjective to almost any noun.

An important tool for recognizing explicitness and robustness problems is a domain model. It allows all activities and conventions within a domain to be understood so that an informations system for the domain provides proper support for these activities and is consistent with the conventions of the domain. Developers of information systems have been constructing domain models for a long time, under various designations, such as conceptual models, enterprise models, or business rules. Software engineers have also become interested in domain analysis and domain models — the different purposes of domain analysis are discussed by Wartik and Prieto-Diaz [9], for a bibliography see [10], Glass and Vessey [11] survey taxonomies of domains.

Consider the domain of software development. A general domain model needs to integrate two kinds of representations of the software development process, representations that deal, respectively, with the managerial and the technical aspects of the process. The most widely known representative of the managerial aspect is the Capability Maturity Model of the SEI [12]. Of technical models there are a good number — some are surveyed in [13]. A managerial model may serve our needs better than a technical model because our aim is not to build software systems but merely to interpret correctly the communications that relate to the building of software.

We now have two contexts — a domain context and a general cultural and linguistic context. The domain context would allow us to understand the specialized vocabulary of the domain. For example, it would suggest that "exploitation system" be replaced by "operating system". How exactly this is to be achieved remains a research topic. One possibility is to build isomorphic domain models in different languages, to construct the intended conceptual model for an application within the domain model relating to one's native language, and then use the isomorphism to create a semantically equivalent conceptual model in English.

6 Conclusions and an agenda for the future

We have outlined some problems faced in international communication. They were grouped under the categories of spelling control, syntactic variations, and context switch. We found that the generation of candidate legitimate words from which to select a replacement for a misspelled word is not difficult. For the other two categories our recommendation is that business and scientific writing for international consumption be explicit and robust. By explicitness we mean that a communication be self-contained, in the sense that no specific cultural frame of reference is needed for its decoding. Robustness means that the meaning of a message is not seriously affected by syntactic variations. The author has observed that Chinese students have great difficulties with English syntax, but that their communications are nevertheless readily understandable. We submit the opinion that this is a consequence of communication by Chinese ideograms being naturally robust. Our first research suggestion is the study of communication in English based on messages composed of root forms of words. These root forms are to be composed similarly to the composition of Chinese ideograms, and a system based on a generative attribute grammar would then convert this structure into text that follows standard English usage. To assist the text generator, the root forms would carry with them linguistic attributes indicating, for example, number and tense.

The second research topic is the study of domain models as a means of clarifying business communications. Just as the initial representation of a message as a composition of English words as quasi-ideograms can assist in dealing with syntax, so domain models can support semantic interpretation. Indeed, any kind of a semantic network is a kind of domain model — some are specialized, such as the domain model for the software industry, others relate to more general semantic relationships of concepts. The domain model could supply the quasi-ideogram with additional attributes.

The domain models lead us back to a consideration of the correction of spelling mistakes. It is easy enough to detect a spelling mistake and to generate candidate corrections for the erroneous word when there is just a single error. What is difficult is the selection of the candidate that fits best into the context. The definition of the appropriate context should be the task of domain analysis, and the selection of the appropriate substitution then becomes a matter of measuring semantic distances in the domain model. Semantic

Proceedings of the 11th International Workshop on Database and Expert Systems Applications (DEXA'00)
0-7695-0680-1/00 $10.00 © 2000 IEEE

networks have, of course, been proposed and implemented before, starting with the pioneering work of Quillian and Sowa, but the merging of the concept of a semantic network with that of a domain model promises a precise partitioning of the concept space into regions, and this, in turn, promises greater precision in finding the right substitution for a misspelled word.

We are resuming work, started in [14], that deals with the use of domain models in the interpretation of queries, and with presenting answers to the queries in context. Consider the query "Was Roosevelt tall?" In answering the query we have three concerns. The first can be formulated as "Who is the Roosevelt of this query?" The second has to do with the fuzziness of tallness, and leads to the reformulation of the original query into the more precise "What is the height of this particular Roosevelt?" The third is more complicated. If the query is finally answered with "The height of Theodore Roosevelt was x feet and y inches," the person who put the query will still not really know whether Theodore Roosevelt was or was not tall. To allow one to make this decision, some comparison value has to be provided as well, e.g., the average height of Theodore Roosevelt's male American contemporaries. The obvious first approach of asking that queries be more specific does not always work, and this is where the approaches discussed above can become relevant. For example, if the person putting the query does not recall the first name of the particular Roosevelt he or she has in mind, something similar to the spelling controller can suggest candidates. This can also help when a non-American cannot recall the exact spelling of Roosevelt.

Acknowledgement

Part of this work was performed while the author was on sabbatical leave in Kaiserslautern. Support was provided by the Fraunhofer-Gesellschaft (Einrichtung Experimentelles Software-Engineering) and the University of Kaiserslautern (Sonderforschungsbereich 501). The support is gratefully acknowledged.

References

[1] Corporate User Publications Group, *Digital Guide to Developing International Software*, Digital Press, 1991.

[2] D. Taylor, *Global Software — Developing Applications for the International Market*, Springer-Verlag, 1992.

[3] T. Madell, C. Parsons, and J. Abegg, *Developing and Localizing International Software*, Prentice-Hall, 1994.

[4] S.M. O'Donnell, *Programming for the World: A Guide to Internationalization*, Prentice-Hall, 1994.

[5] C. Zhang and R.F. Walters, "An abstract, shared and persistent data structure for supporting database management and multilingual natural language processing," *Int. J. Software Eng. Knowledge Eng.*, Vol. 3, pp. 369–382, 1993.

[6] C. Paris and K.V. Linden, "An interactive support tool for writing multilingual manuals," *Computer*, Vol. 29, No.7, pp. 45–56, July 1996.

[7] A.T. Berztiss, *Software Methods for Business Reengineering*, Springer-Verlag, 1995.

[8] K. Nakakoji, "Beyond language translation: crossing the cultural divide," *IEEE Software*, Vol. 13, No. 6, pp. 43–46, Nov. 1996.

[9] S. Wartik and R. Prieto-Diaz, "Criteria for comparing reuse-oriented domain analysis approaches," *Int. J. Software Eng. Knowledge Eng.*, Vol. 2, pp. 403–431, 1992.

[10] W.A. Rolling, "A preliminary annotated bibliography on domain engineering," *Software Engineering Notes*, Vol. 19, No. 3, pp. 82-84, July 1994.

[11] R.L. Glass and I. Vessey, "Contemporary application-domain taxonomies," *IEEE Software*, Vol. 12, No. 4, pp. 63–76, July 1995.

[12] M.C. Paulk *et al*, *The Capability Maturity Model: Guidelines for Improving the Software Process*, Addison-Wesley, 1995.

[13] P. Armenise, S. Bandinelli, C. Ghezzi, and A. Morzenti, "A survey and assessment of software process representation formalisms," *Int. J. Software Eng. Knowledge Eng.*, Vol. 3, pp. 401–426, 1993.

[14] A.T. Berztiss, "Imprecise queries and the quality of conceptual models," in *Information Modelling and Knowledge Bases V*, pp. 174–185, IOS Press, 1994.