Ethical Dilemmas for the Computational Linguist in the Business World

Heather McCallum-Bayliss
Language Analysis Systems, Inc.

Most applications of computational linguistics take place in commercial settings, where issues of concern to the businessperson may be of less interest to the linguist. Commonly, the linguist struggles with contributing worthwhile linguistic insights and working to meet the commercial ends of the client.

This struggle often emanates from the differing worldviews held by the linguist and the business professional. This contrast is especially noteworthy with projects that involve multiethnic/ multicultural/multilingual databases. Predictably, such projects necessitate a heterogeneous view of the data, one that is less amenable to the structure and predictability generally associated with computers. The project that will serve as the basis for most of the discussion in this article deals with the field of onomastics, a field filled with variation. Personal names can vary in many ways: For example, (1) in name structure (Hispanic names have two surnames, the first of which is the "last name;" Brazilian names generally have two surnames, the second of which is the "last name;" many cultures have only one name); (2) socially (names may change on marriage; special markers may be added as social status changes); (3) in predictable spelling variation (Barton and Varton would be considered two different name types in the Anglo naming system; but Bargas and Vargas could be variants of the same name type in the Hispanic naming system); (4) in the character set or writing system used (Spanish uses accents and a tilde, e.g., Cánepa, Muñoz; Chinese, Arabic and Russian have unique writing systems and transliteration issues arise); and (5) in name frequency (Garcia is a highly frequent Hispanic surname; Kim is highly frequent in Korean names).

Issues in Applied Linguistics ISSN 1050-4273

Vol. 4 No. 2 1993 257-269

In the project under discussion, the client needs to retrieve records from a 20-million-record multiethnic/multicultural/multilingual database of personal names. The records retrieved from the database should contain the exact name requested or a close approximation to it. The client had found that the existing, essentially Anglo-centric system employed for this task was unreliable and inconsistent in its retrieval of a record with an exact name match, and that it produced intuitively unsatisfactory results when attempting to return records with close approximations to the name requested. The inconsistencies in the retrieval results stemmed both from problems with the nature of the computational algorithm being used and from the failure of the system design to recognize that personal naming systems vary from one cultural, ethnic and linguistic group to another. An Anglo- (or Euro-) centric view of names could not produce adequate results in a multicultural/multiethnic/multilinguistic database.

The role of the linguist in this project was to attempt to make the system more sensitive to the cultural variation inherent in onomastic systems. This role exposed many areas of tension between the goals of the linguist and those of the business client, which led to questions about what constitutes appropriate behavior for the linguist. I will explore six of these areas of conflict and discuss how each may give rise to an ethical dilemma for the linguist:

1. A conflict between the client's desire for confidentiality and the linguist's desire to share linguistic data and research findings with colleagues.
2. A conflict between the business orientation of the client and the theoretical orientation of the linguist.
3. A conflict between established project parameters and the accommodation of linguistic insights within these parameters.
4. A conflict regarding ownership of products that result from application of linguistic insights.
5. A conflict between the client's limited understanding of the linguistic implications of the research and the linguist's responsibility to educate the client and/or user.
6. A conflict between the client's and the linguist's responsibility for the success or failure of the product.

ETHICAL CONSIDERATIONS

Confidentiality vs. Dissemination

Many of the problems encountered in the project cannot be reported directly since the client has asked that the nature of the tasks, the goals of the project, and examples from the database not be divulged. Most of the examples in this discussion, therefore, correspond to data in the project but are not taken from the data directly.

Requests for client confidentiality present the linguist with a predicament. Maintaining confidentiality makes it extremely difficult to share linguistic insights with colleagues. When the material on which linguistic generalizations are based cannot be subjected to public scrutiny, a basic tenet of linguistic research is contravened. However, the linguist is being paid to work on a particular project. What sort of obligation does such payment entail? Does the linguist have to adhere to conditions that impede the advancement of scholarship, or could it be argued that submitting material to the linguistic community does not actually breach a client's confidentiality since it is not in the business setting that the information is being divulged? The linguist must also face the issue of whether or not it is sufficient to provide the linguistic community with examples similar to those found in the database, if the linguistic generalizations are not affected. After all, "concocted" examples have long been used by syntacticians, if they correspond to native speaker intuitions.

The need for confidentiality stems from regulations or industry conventions with which the linguist may not be fully familiar. She must therefore accept and respect the client's knowledge about the consequences of breaching confidentiality. To ensure that some, if not all, aspects of research can be shared with colleagues, she must be certain that she understands the client's notion of confidentiality and that she is able to abide by the restriction it poses. In the present case, most examples correspond to data in the project, but they are not taken from the data directly.

An interesting corollary to the issue of confidentiality was mentioned by a reviewer of an earlier draft of this paper. Does the use of examples parallel to those found in the data breach the intent, if not the letter, of the confidentiality condition? This is an

interesting problem, and one that I believe can be answered relatively straightforwardly. Automatic namesearching is a relatively new undertaking, but the principles involved are for the most part consistent across projects. The linguistic generalizations and representative examples will apply to any project whose database is ethnically/culturally/linguistically diverse. The goals of the projects may vary; the data may vary; the motivations for confidentiality may be different; the systems may have different designs; but the underlying principles of namesearching will be the same. So, use of analogous examples does not violate the client's request for confidentiality, since the generalizations stem from general linguistic research and not from specific examples found in the database. Generalizations based on a specific database may need to be made more generic when they are reported, so as to mask the client's specific purpose.

Another dilemma concerning confidentiality may also arise. How far does the obligation of confidentiality range? What if, once the project begins, the linguist finds that the linguistic knowledge that she is supplying is being utilized in ways that she does not condone. To whom does she owe her allegiance: to the client, who is paying her; or to the public, who she believes has a right to know the nature of the project? One would hope that the linguist would not accept work on a project whose purpose she found morally repugnant. The contract process is likely to provide a sufficient amount of detail about the nature of the project to deter involvement in a project that seems objectionable. However, project aims should be carefully scrutinized during the contracting process.

Business vs. Theory

While the need to be responsive to cultural variation may seem obvious to the linguist, it may not seem relevant to the linguistically naive client. This dichotomy raises the broader conflict between the business orientation of the client and the theoretical orientation of the linguist. She may need to spend a significant amount of time educating the client about relevant linguistic issues and convincing the client that the linguistic generalizations presented are crucial to solving the problem at hand.

In the project under discussion, cultural variation is evidenced orthographically. In the Anglo-American culture, Wiley might be spelling Wilee or Wily or Weiley; or Brown might be

found as Browne or Braun. However, Boland will not be spelled with an initial V (Voland) and be considered a variant of the same surname; it will be a different surname altogether. Hispanic surnames, on the other hand, *do* have surnames in which B and V alternate: Bernal and Vernal. But just as the exchange of B and V will produce different Anglo name types, the substitution of RR for R will produce two distinctive Hispanic surname types (Moro and Morro). (Note that variants with R and RR [Ferris and Feris] in Anglo names produce spelling variants only.) With examples such as these, it becomes quite clear that spelling variants which are peculiar to a specific linguistic system have implications for system design. If a computational system has the mandate to retrieve records that approximate a name submitted, then the system must be made to recognize the linguistic variants that are applicable in a cultural group.

Convincing the client that spelling variation is language-specific may not be difficult; more challenging will be arguing that these linguistic generalizations should be incorporated into the algorithm. While the issue for the linguist is how to include a linguistic generalization, the problem for the client may be one of programming efficiency. The argument could proceed along the following lines.

The client may understand that there are spelling variants in Hispanic surnames that do not exist in Anglo surnames and that the existing system has not been sensitive to these distinctions. However, he may propose that instead of redoing the algorithm completely to incorporate these variations, which would cost time and money, a list of the alternate spellings for the names be incorporated into the system, since computers are especially efficient at comparing items to a list. Each name would be listed separately and the algorithm would be made to proceed as follows: If the surname Sanchez is being sought in the database, the search should also include any name that is spelled Sanches.

While the linguist would concede that computers are good at going through lists, she would have to argue that listing alternate spellings of names introduces problems. First, it misses the linguistic generalization that B and V alternate in Hispanic surnames. Secondly, and more importantly, it presumes that every name variant can be anticipated and put on a list. It would be far more efficient to develop a rule that states that S and Z alternate in particular environments in Hispanic surnames.

At this point, the client and linguist are at somewhat of an impasse. To incorporate the generalization, the computer would have to examine each letter in a name, which would utilize a significant amount of processing time. More problematically, since this spelling rule is language/culture specific, the linguistic generalization presumes that the computer can recognize that a surname is Hispanic. How can the B/V rule be constrained to apply only to Hispanic names and not to the names of other cultures? If other relevant information is available, such as country of birth, it could be utilized to help constrain the rule. Even that information is suspect, however. Countries are not homogeneous cultural groups; many non-Hispanics can be found in predominantly Hispanic countries, and many Hispanics can be found in predominantly non-Hispanic countries.

With these complications, the client may be reluctant to value the linguistic scholarship offered, regarding it as overly theoretical, esoteric and not compatible with commercial ends. He may consider his needs minimally linguistic and more a question of programming efficiency. And he may be disinclined to restructure a system entirely to accommodate a world-view that seemingly can be achieved through programming changes. He may conclude that including linguistic insights would necessitate an entire rethinking of the system design, since a linguistic perspective on a project may cause a different way of viewing and handling the problems encountered. Remodeling the system may lead him to disregard linguistics altogether.

For the linguist, there are questions as well. Since the insights presented about spelling variation would necessitate significant restructuring of the system and a likely increase in processing time, she must weigh whether or not such a change is warranted by the frequency and import of the problem. That is, she must determine how likely it is that spelling variations could occur and how damaging the consequences would be if the less frequent spelling variations were not included. She is faced with the conflict of respecting both the commercial concerns of the client and the accuracy and linguistic adequacy of the system.

Other material concerns may also enter the picture. Time and budgetary considerations and the recent explosion in the volume of electronic data storage may constrain the ability to draw adequate generalizations from the data. The linguist may be forced to compromise the depth of analysis and research in order to meet

market deadlines and to curb the costs of preparatory work. For example, assume that the system is returning multiple examples of inadequate matches, such as:

REQUEST:	JORGE LUIS SANCHEZ RODRIGUEZ
RESPONSES:	JORGE LUIS SANCHEZ LOPEZ
	JORGE LUIS SANCHEZ SAAVEDRA
	JORGE LUIS SANCHEZ BUSTAMANTE

These examples contain non-matching matronymics (the second surname), which makes them unlikely candidates for a match with the name requested. If the addresses of these individuals were available, the system could be designed to examine a reduced number of records based on smaller geo-political entities (e.g., counties, territories). This would certainly accomplish the goal of limiting the number of poor responses, since the number of records examined would be fewer. However, by no means would it address the underlying problem of non-matching matronymics: A solution based on geography merely avoids the crucial linguistic issue.

The inappropriateness of this solution is relatively transparent. Other cases are often less clear cut. Confronted with the need to compromise her analysis to meet a client's deadline, the linguist is then faced with the question of what such compromises do to the adequacy of the analysis and whether or not what she is offering are actual linguistic insights or merely "fix-its" that mask an underlying problem. She must constantly balance the magnitude and import of a problem and its linguistic solution with the impact that the solution will have on the cost of design and speed of operation of the system. She must be sensitive to the client's requirements and the goals of the system. Some issues will be worthy of altering in the system and some will not. The linguist must find the balance among all these considerations. The dilemma for the linguist results from the need to adhere to high standards in solving linguistic issues but at the same time honoring the commercial needs of the client.

Conforming Linguistic Insights to Project Parameters

Consultants may face pre-determined constraints that limit the scope of possible scholarship. For example, the system may have been designed to use a particular programming language.

Certain programming languages perform certain functions better than others: some manage large quantities of data while others are able to manipulate small units or subsets of data; some deal well with natural language while others deal especially well with numbers. The pre-existing conditions that the linguist finds on entering a project can preclude application of the most appropriate and relevant linguistic processes.

For example, if the system can handle only certain sizes of arrays, then comparison of particular types of personal names may be limited. Arabic names, which tend to permit much variation in the number and ordering of name elements, present a problem here. A single name may have the following variants:[1]

ABD RAHMAN MUHAMMAD ABU AL MAJD
MUHAMMAD ABD AL RAHMAN MAJD
MOHAMMAD AL MAJD
MOHAMMAD ABU AL MAJD ABD AL RAHMAN

A system would have to be able to compare virtually any name element with any other name element in order to determine if a match could be found. This process would require an algorithm that could manipulate and compare many elements and store information on the likelihood of a match. If the processing language in use is one that is not capable of such tasks, the responsibility arises of proposing solutions that are less than optimal given the constraints of the software that was chosen before linguistic issues were presented to the client. On the other hand, if the linguist insists on changing the project resources, she may run the risk of having all linguistic information rejected: Some linguistic sensitivity in a project is likely to be far better than none at all.

Issues of Ownership

Because linguistics in the commercial realm is relatively new, it is not uncommon for a project to spawn commercially viable, linguistically based products. To whom do these products belong? If they result from the work performed by linguists and have other linguistic applications, do they belong to the linguists or to the client? If, for example, the computational linguist develops and encodes a namecheck system that can examine and retrieve names from a large multilingual/multicultural database quickly and

accurately, does that system belong exclusively to the client? It could be argued that the client has one goal in mind—to accomplish the task that is his mandate. He may not have the time to produce a marketable product or the interest. If that is the case, then further applications of the product may be in the hands of the linguist to pursue. Such contingencies must be explicitly prepared for when contract negotiations are underway. However, even if the question of ownership of commercially viable products is anticipated, ethical questions about how much of a product belongs to a client can still arise. If a product addresses the needs of the client but is limited in its scope, do all extensions of the product also belong to the client or to the persons who pursue the further research and development of it?

Another question arises in this regard: Is it the linguist's responsibility to educate the client sufficiently to understand the detailed operation and potential applications of a product so that the client could make use of it in other settings? Does a product that can be altered or improved (perhaps generalized) to fit other commercial endeavors still belong to the original employer or to the developer, since it is not *exactly* the same product?

The answers to these questions are not clear. Unquestionably, extraordinary care must be exercised to specify from the beginning what it is that will be done for the client and what the disposition of both planned and unanticipated products is to be.

Educating the Client

Generally, commercial enterprises focus on producing a market-ready product as quickly as possible. Training for use of such products rarely goes beyond operational details and may not include material on what the system is capable of doing or on the application of the product to other existing or new linguistic problems.

Training manuals for a namecheck system must include information on entry format, discussion of error messages, and details on how to manage data in the system, but if they do not include information about the linguistic assumptions that have been made about the names returned, problems are likely to arise. For example, if the user is unfamiliar with Hispanic names, he may be

surprised to find that the following are appropriate results for the name requested:

REQUEST:	JOSE LUIS DELACRUZ
RESPONSES:	JOSE LUIS DELACRUZ RIOS
	JOSE DE LA CRUZ RIOS
	JOSE DELACRUZ

Since it cannot be assumed that every user is familiar with every type of cultural variation in names, a resource manual is crucial to assist the client in understanding what assumptions have been made about names in the system that he is using. In most cases, the scope of a particular namecheck system is limited; due to constraints of time, budget, and purpose, not every possible name variant can be anticipated. Knowing which sorts of variants have been included is therefore helpful to the user. For example, can the user expect to find spelling variants such as Bargas and Vargas, or not? Such information will keep the user from becoming frustrated with the sorts of responses that the system may deliver.

Similarly, the user would need information about name variants that one might try entering if certain types of results are desired. For example, if the name of a married Hispanic woman, Luz Carmen León De García, does not produce the desired responses from the system, it may be that the individual's name is in the system in maiden name form, Luz Carmen León Orozco. It would be possible to try to find a match by reentering the name minus the marriage marker De and married surname, Luz Carmen León . If the user is unfamiliar with the nuances of a culture's personal naming system, then such alternate attempts at a match would not be possible. A training manual would be useful in this regard.

Finally, a manual would need to include information regarding the consequences of inconsistent data entry practices. Although data entry format guidelines may be specified, the motivation behind such guidelines and the consequences of not following the instructions may not be at all clear. If one user among many assumes, for example, that all Hispanic surnames should be "Europeanized" and the patronymic and matronymic surnames should be inverted (Gomez Torres would become Torres Gomez), the crucial surname (patronymic) information would have shifted from one position to another. If the design of the system has not

anticipated such a change, then requests for this name from other users would produce no results. Again, a reference manual would make clear the assumptions about the naming conventions of the various cultures included.

It is clear that training must be a central concern for the linguist. She is offering the client new ways of addressing data management problems and providing new resources for the resolution of these problems by broadening the client's views and assumptions about multicultural/multiethnic/multilingual environments. For the client to extract the full value from the system and to avoid misuse of the system, he must be adequately trained. The linguist has a responsibility to educate the client; it is not sufficient to provide only the linguistic insights that make the project work more efficiently. The scope and definition of the educational effort poses a challenge for the linguist.

Responsibility for Project Success and Failure

If difficulties arise in the operation or function of the product or system being produced (such as system overload), who bears the responsibility for failure? What does such responsibility entail? Several of the examples discussed above involved solutions that would have increased processing time and energy significantly. Such solutions not only affect the cost and time required to produce the system but may have a marked impact on the processing efficiency of the product. If the response time to the user is a crucial issue and the linguistic processing that must be done slows the system or even overloads it, the linguist may have maintained her linguistic integrity but at the cost of interfering with one of the central goals of the project. Is system failure the responsibility of the consultant who introduced more adequate and accurate, yet slower, name matching techniques into the system? Or is it the responsibility of the client, the programmer, or other participant, who may not have provided adequate programming techniques or may have restricted the range of possible linguistic functions through choice of particular programming language, etc.? It seems clear that the linguist would have to bear much of the responsibility for project failure if she were to propose linguistic solutions that would have deleterious effects on the general operation of the

system. She must, therefore, be mindful of the abilities and needs of the client and of the goals of the overall system.

CONCLUSION

It is clear that the computational linguist working in the marketplace frequently encounters the tension between the client's commercial goals and her responsibility to provide adequate, appropriate scholarship. The examples cited above are not exceptional. These difficulties recur frequently and in many different guises, and although the discussion has focused on one project, the issues raised apply to virtually any computational project anchored in the business community.

The conflicts the linguist encounters give rise to three types of responsibility: 1) social responsibility; 2) ethical or moral responsibility; and 3) professional responsibility. She has a social responsibility to educate, to provide enough information to the client so that he will understand the importance of a culturally/linguistically/ethnically sensitive view of the world and the data. Such information will not only make the project more successful but will make clear the need for linguistic scholarship in the marketplace.

The linguist has ethical or moral responsibilities as well. She may struggle with honoring a client's request for confidentiality and her own misgivings about the nature of the project. Or she may contend with the issue of determining how much of her linguistic knowledge and project expertise is the property of the client. What of this knowledge can she ethically use in other endeavors without obligation to the original client?

Finally, the linguist has a professional responsibility. She must maintain her integrity in providing suitable linguistic insights to the client and must strive to share linguistic generalizations from the project with colleagues. Pre-determined project conditions test the linguist's ability to provide adequate linguistic detail and, as a consequence, limit the insights she may glean from the data. Is there an obligation to provide all the linguistic information that can be identified and, as a result, have abundant insights to share with colleagues, or is there an obligation to contribute only those insights

that will lead to the successful completion of the project, perhaps limiting the value of the conclusions to the field?

Quality control is another professional responsibility. The "scientist" will undoubtedly want to define *quality* as 'accuracy' and 'inclusivity,' but 'consistency' is may be a better term in the commercial domain. Consistent, predictable results promote confidence in the user. The linguist must be dedicated to linguistic principles, to discovering and arguing for those generalizations that are crucial to the definition of the cultural, ethnic, linguistic variation but she must also be able to work within the defined parameters of the commercial project.

The business world demands compromise of the linguist. The compromises must be weighed against the social, ethical and professional responsibilities that she has. Finding the delicate balance between the theoretical and practical is a recurring battle for the linguist in the business world.

NOTES

[1]Thanks to Dr. Karin Ryding for providing these examples of Arabic names.

Heather McCallum-Bayliss is a senior computational linguist with Language Analysis Systems, Inc., of Chantilly, Virginia, a firm specializing in the development of namesearching systems, natural language processing, and computer-assisted language learning. She received a Ph.D. in theoretical linguistics from Georgetown University in 1984. In addition to computational linguistics and computer system design, her work has included language pedagogy, lexicology, and computer-assisted language learning.