How the study of endangered languages will revolutionize linguistics

D. H. Whalen
Haskins Laboratories and the Endangered Language Fund

The field of linguistics is heading into a time a great change. The fundamental elements assumed by the field are undergoing radical revision, fueled by two factors: the computerization of language material, especially from endangered languages, and the more detailed study of endangered languages themselves. The improvement in our ability to document languages is occurring even as the languages cease to be spoken, increasing the urgency of the documentation process. Although documentation is essential to revival efforts and thus of great use to the native communities, this paper focuses on the dramatic influence such language material will have on the field of linguistics.

1. Introduction

Linguistics is at an exciting stage of development, poised to make more dramatic discoveries than at any time in the past 40 years. Unlike the conceptual advances that were largely responsible for the Chomskyan revolution (e.g., Harris 1993), the changes today are fueled by changes in the nature of the data and the technology that deals with it. The most fundamental elements of the field of linguistics are being revised, and the study of endangered languages will be at the center of this new revolution.

Let me hasten to add that the study of endangered languages is not new to linguistics – far from it. Many languages are now only known through the recording efforts of earlier linguists. Indeed, it is through these archives that the languages can be called "sleeping" rather than dead – various Native American groups have used this term, as outlined in Hinton (2001). Even at the formation of the Linguistic Society of America, Leonard Bloomfield noted the state of the "American Indian languages, which are disappearing forever, more

rapidly than they can be recorded, what with the almost total lack of funds and organization" (Bloomfield 1925:4). In the 75 years since then, several things have happened: The rate of language death has increased, but the means of recording languages has expanded dramatically. The rate at which languages are "going to sleep" is higher, but the availability of materials from which we might awaken them has increased as well. Oddly, it took another 65 years for organizations devoted to stopping language loss to appear, and the money has begun to flow in sizable amounts only in the last few years. The damage done to societies when they are forced to give up their language have been detailed elsewhere (Crystal 2000; Grenoble & Whaley 1998; Hinton & Hale 2001), so I will not go over the details here. Suffice it to say that language loss in the modern world seldom reflects a voluntary choice on the part of the language community.

We have already heard from Tasaku Tsunoda (this volume) about the effective use of linguistic materials, collected for purposes of analysis, but now the only record of a language. The fact that we can revive a version of a language that is satisfying to the heritage group (that is, those descended from the original speakers) is a tribute to the level of accomplishment we have attained in the field of linguistics. The use of material by native speakers and their descendants is an extremely important aspect of dealing with the modern world in an ethical manner. Although the ethical use of language is a complicated topic, it is one that is receiving a much needed upsurge in attention (e.g., Hinton & Hale 2001; Maffi 2000, 2001; Whalen 2001).

1.1 Focus on the benefits to linguistics

My focus today, though, is the effect of the study of endangered languages on the field of linguistics itself. Thousands of languages are in danger of disappearing without, in many cases, much of a record left behind. The fact that they are ripe for study is no guarantee that recording them would revolutionize the field of linguistics. The study of vanishing species does not necessarily hold the same promise for biology, for example. If the causes of speciation and genetic inheritance are the same in the extant species as in the endangered species, then biology might have only incremental lessons to learn before the species disappear. It is certainly true that species are being examined for unusual chemical products that they might produce, especially ones with medicinal value to humans. But the fundamentals of biology seem secure. Linguistics, on the other hand, has had some success at describing its subject matter, but there is no

grammar that is a complete and satisfactory one. The lessons we learn from endangered languages will change that within the coming decades.

The concept of language documentation has also been a rather recent addition to the discussion in linguistics. Christian Lehmann (personal communication), writes, "The primary purpose of language documentation is to represent the language for those who do not have direct access to the language itself." Thus the text collection that has proceeded throughout the history of our field is only a part of documentation. No text fully represents a language, and texts must always be supplemented with analysis if the language is to be known in anything like its full form. Documentation is not critical for the "safe" languages, the ones that will continue to be spoken, yet that is where most of the effort is concentrated. For example, in the excellent collection of material at the Linguistic Data Consortium (www.ldc.upenn.edu), over half of the data sets are for English. Among the rest, only Ngomba is a small language, the next smallest being Czech and Farsi. This situation needs to change, and there are indications that it will.

1.2 Two main aspects of the coming changes

There are actually two parts in the realm of changes to linguistics that will take place, and only the second is truly unique to endangered languages. I will get to that later, but let me first discuss an aspect of language analysis that is becoming common to all language study and that will have revolutionary impact: the computerization of speech. Although this in itself is not new, the effects have yet to be felt. The Association for Computers and the Humanities, for example, was founded in 1978, to promote the use of computing in fields like linguistics. There are many fine programs that do such interesting tasks as constructing syntactic trees, watching the descent of daughter forms from proto forms, making dictionary entries accessible electronically, etc. But the impact has been limited so far. What needed to happen was to combine the power of the computer itself with the universal access associated with the internet. Now that such a union has occurred, there still need to be some changes before the revolutionary impact can be seen.

Having a computerized text is somewhat more useful than having a transcription, but it is not revolutionary. Being able to search through the text and have a program perform statistics on the distribution of various linguistic features is quite useful and can lead to insights that would not be feasible otherwise. The study of contested authorship, the frequency of occurrence of phonemes and words, and the acoustic analysis of the speech signal would all

be forbiddingly difficult without a tool like the computer. But the way that digitized language has been shared to date is much more like the longest tradition of linguistic field work, in which notes about the utterances reside with a single researcher and are available only to those who physically travel to where the notes are and receive permission to use them.

What is beginning to happen now is completely different, namely, original linguistic material is being made available to the broadest possible audience by putting it on the internet. There are four reasons this will change the way linguistics is done:

1. Linguists will be working from the same data rather than individual data sets.
2. The texts can be searched easily in many ways.
3. The actual sounds will be available rather than just transcriptions or further idealizations.
4. Work will progress on a unified ontology for linguistics.

The first two issues are quite dependent on the technological advances provided by the internet. It might seem that the fourth, the development of a single set of descriptive terms for linguistic structures, could happen regardless of changes in technology, but I will try to show that this is not the case.

2.

2.1 Access to data

The sharing of data is critical to advancing the field. Imagine the state that the Neo-Grammarians would have been in had they not all had access to the same data. Imagine that Bopp had the Rig Veda and Whitney had the Atharva Veda, and they were the only ones that could see them. We would never have had the explosion of progress in comparative analysis that we did see. Instead, the texts were widely available and widely used. Granted, they were not best-sellers even then, but libraries throughout the world could provide the raw material for linguistic analysis. If anyone felt that a citation was incorrect, she could look it up. It seems uncontroversial that this is a good way of doing science.

But compare that with the situation that we are in now. For common languages, we have a wealth of material along with native speaker intuitions. But for looking at more unusual languages, we are at a stage where essentially one or two linguists have access to the raw data, and everyone else gets snippets of

their analysis. Even in the best case, in which a linguist has deposited recordings with an accessible archive like the American Philosophical Society or Canada's Museum of Civilization, access is limited to those that can make a trip to those locations. Further, the amount of work that is needed to extract material from a tape recording that is deposited in an archive is extremely large. In the best case, there is a transcription of the entire tape to start from, but aligning that transcription with the tape itself is a technological challenge. In many cases, there is no transcription but only a translation, which requires even more expertise on the part of the would-be user. It appears to be even more common for large sets of materials to end up in archives with no accompanying interpretation at all. At this point, one must almost be a native speaker in order to make use of the material.

In order to make the basic facts of a wide range of languages available to the broadest range of linguists, we need the internet. Publishing texts is an improvement over the more typical state of affairs, but print publishers are increasingly reluctant to produce such volumes. They are typically not profitable and end up with limited distribution. Having a published version that very few people have access to is only a slight step ahead of not having a published version at all. With the internet, the initial costs are all that have to be considered. Once a text is put on the web, it can be accessed once or hundreds of times, and the resulting costs are essentially the same. This allows for wide dissemination without excessive cost.

Incidentally, this version of publishing is increasingly useful for native speaker and heritage communities as well. More and more groups are gaining access to the internet, often before they have many other services. In some cases, they have internet access even in refugee camps, allowing access that would not be possible with printed material.

2.2 Searching the data

The second important aspect of putting language material on the internet is the searching capabilities. These are very useful for the kinds of research that linguists do, and they allow research that would be all but impossible with printed texts. Imagine searching through texts for every occurrence of the phoneme /p/, and only deciding at the end of the search that you also needed every instance of /b/. Hundreds of hours of work, going over the same material, would ensue. With an on-line search, much of the work is done even before the search is attempted. Words must be tagged separately, and each phoneme in a fully marked-up text will already be accessible. The additional work needed to find

and count different classes of phonemes is then relatively trivial. Certainly the actual searching is done by a machine that does not really mind being asked to do the same thing over and over, with slight variations. After all, it's just more electrons flowing through as far as the computer is concerned. For the linguist, it is the difference between having results and having none. Electronic publishing of texts and other linguistic material, therefore, will bring us back to the stage we were at a hundred years ago, when any interested linguist could go to the original source material if she so desired.

2.3 Access to sound files

The third aspect of electronic publishing, however, is even more revolutionary: the inclusion of sound files. As linguists, we know that orthographies record as much of a language as native speakers need to easily reconstruct an utterance, and thus are of varying use for linguistic analysis. We also know that transcriptions, even the most narrow ones, are an idealization of the signal. Yet we have been limited in our reanalysis of other researchers' materials to just these compromises. Having a sound file available with the text is a way to overcome these limitations to a large extent. Recordings themselves are something of a compromise, since visual information is lost and aspects of the recording situation can affect the apparent linguistic sounds that are recorded. As a simple example, if the recording situation was in someone's kitchen, it may be impossible to tell whether a certain syllable contained a click or was accompanied by a can being opened in the background. However, the amount of phonetic detail that is captured is several orders of magnitude larger than that in the transcription. This information can now be shared at the click of a button. An excellent example is the work on tone in Bamileke Dschang (Bird 1999), with a web version available at http://www.ldc.upenn.edu/sb/fieldwork/. Here, the rather subtle differences between the tones can be heard and other, related forms can be found as well. As the linked texts become more common, even more examples of the tone contrasts would become available to linguists all over the world.

Phonology and phonetics certainly stand to gain from the inclusion of sound files. It is often the case that a linguist will question data in published reports on a language that she knows well. With the sound file included, it is possible for any linguist to make a judgment. Caution is certainly necessary; it is usually the case that subtle aspects of an utterance will elude the untrained ear. Yet it is also the case that four ears are better than two, at least as a corollary to the saying, "Two heads are better than one." One listener will often detect

something that another listener will miss, and sometimes being unfamiliar with the language is a positive benefit. It should certainly be possible, in any case, to convince listeners that the expert's analysis is correct – the subtleties should ultimately be perceivable.

Syntax, though, may benefit even more than phonology and phonetics. Our current transcription systems are rather schematic, as already mentioned, but they are typically ignored altogether for intonation. The use of intonation to signal syntactic relations is one of the hallmarks of language, yet most syntactic analysis proceeds without regard to intonation. This is largely a practical decision, though it is sometimes raised to the status of a theoretical one (e.g., Martinet 1960). What is quite clear is that certain sentences that are thought to be ambiguous when examined in written form are perfectly unambiguous when spoken. As a headline, "British left waffles on Falkland Islands" (Lederer 1989:63) is ambiguous. When spoken, it would usually take a fairly uncooperative listener to take the wrong meaning. This is a simplification as well, since there are intonations that are rather ambiguous for this sentence, but we have not yet explored how frequent such constructions are nor how often speakers use them in favor of an unambiguous intonation. The data has, for the most part, been unavailable. With immediate access to spoken versions, we will be able to assess the intonation more directly. The concentrated attention of large numbers of syntacticians may also lead to a more useful narrow transcription system for intonation.

It is worth mentioning here that we will probably come to expect that linguistic examples in published papers will have spoken versions available on the web. This will be of great assistance to those who are reading the articles in evaluating the data. Often times, the reader is not a native speaker of the language in question, and so may be unaware of the kind of interpretation it would be given. Even for those who speak the language, dialect differences often are not apparent in the written form, or even the transcription. Ultimately, this will allow a more complete description of the language data to inform the theoretical debates.

2.4 A unified ontology

The fourth improvement in linguistic analysis that will be due to putting material on the internet is the development and use of a unified linguistic ontology. An ontology in the modern, nonphilosophical sense, is a description of objects and their relationships to each other. While the description of ontology might seem like an atheoretical task, it actually goes to the heart of a discipline, es-

pecially one like linguistics in which the "objects" are all mental categories and thus difficult to find evidence for. The use of a unified ontology will be felt both because the machines require it and because it is the right time to tackle this problem. The current state of linguistic theory contains an extreme contradiction in what we think about our subject matter. On the one hand, we assume that all languages are basically the same. Every child seems to be able to learn any language (though I will return to this later), and the basic building blocks seem to be the same for all languages. However, all languages are different. This is true first in the trivial sense that if they were not different, we would not call them separate languages, but it is also true that they differ in ways that are not easy to describe as just being minor variations of universal categories. Yet how can this be, if all languages are the same? It is always possible to take an extremely general view of language and say that the really important part is only a single aspect such as recursion (Hauser et al. 2002), but when the grammars are written, this level of analysis is insufficient to capture the patterns of language that are found. The descriptive vocabulary in particular seems to fail us as we go from one language family to another (and sometimes from one language to another).

Let's take an example from a universal in human languages, the treatment of aspect. Every language has some way of expressing the relationship of events to a time-line. Sasse (2002) lists three areas of general agreement on this issue: aspect is a matter of boundaries; it is governed both by the grammar and the interaction of the grammar and the lexicon; and explicit theoretical descriptions of syntax, the lexicon and their interaction are needed for an understanding of aspect. Sasse then states, "Within these general confines, there are contradictory positions on almost any of the basic issues: there are considerable differences in the various underlying models of 'time logic'; there are divergent views on what constitutes 'atoms' of an aspect theory (i.e., the granularity of distinctions); there is disagreement on the interpretation of the different semantic effects observed (truth-conditional values, pragmatic implications, universal semantic categories, etc.)" (2002:202). Before discussing more specific issues, it is startling to see that such a universal feature of language has not allowed for a universal and agreed upon analysis. This tells us that simply being a universal part of language is no guarantee that a unified description is obvious and adequate. The disagreements about the terminology come about not because the temporal domain is so complex but because the interacting domains that different languages take to be relevant are numerous and idiosyncratic. Sasse concludes that much has been accomplished, but that individual languages must be examined to confirm or refute our cur-

rent theories. The internet will make that more feasible, since examples that have been labeled with one category or another can be easily compared and the underlying data examined by all interested researchers.

Sasse also recommends an "open-minded research strategy" (2002:266), which appears to be at odds with the notion of creating a single ontology for linguistics. However, the two must go together. If we do not attempt a unified ontology, then it will be impossible to make use of the rich source of data that the internet is beginning to provide. Only when the disagreements about labels can be argued on the basis of data rather than definitions can we begin to have a true science of language. To take an analogy with chemistry, we have not yet discovered the periodic table of linguistics. We do not yet know what evidence it would take to establish a new "element," let's say, a new morphological category. But surely if we do not attempt to restrict the elements, then we will never know which ones are truly needed. The chemical analogy goes further, in that elements interact with other chemicals differently in different ionization states. Once these states are recognized, the fact that elements behave differently in different environments is predictable rather than a cause for disagreement. We are only beginning to find such a description for linguistics.

Using a single ontology will require changes in areal grammatical terms, and this will cause disruption in the short term that should be justified in the long term. Consider the "obviative" marking in the Algonquian languages. This complex ranking of third person cases is a hallmark of all the Algonquian languages (Bloomfield 1962; Frantz 1966; Hockett 1966). Yet is it a category that does not exist in any other language family? Should we expect to find that the existence of entire morphological categories is a feature of being a language family? Recent work by Aissen (1997) indicates that we should expect other languages to have processes similar enough to obviation to deserve being classified as such. She examines two languages unrelated to Algonquian – Tzotzil, a Mayan language, and Chamorro, a Western Austronesian language – and finds that similar processes are at work there. It is only through the unwillingness to take a language family label as the ultimate best description that allows us to uncover similar patterns, and presumably similar solutions, in other languages. Although Aissen's work shows that this kind of effort can be accomplished without the impetus of a unified ontology, the ontology will spur many more researchers on to this important kind of work.

One of the most extensive attempts at a linguistic ontology is the "General Ontology for Linguistic Description" (GOLD) being developed by Terry Langendoen, Scott Farrar, and William D. Lewis. They are working under the EMELD project, funded by the U.S. National Science Foundation

(www.emeld.org). This five year project is devoted to solidifying the techniques and metadata needed for presenting endangered language material on the web. It is under the direction of Helen Aristar Dry and Anthony Aristar, whom most of you will recognize as the editors of the Linguist List. Not having been satisfied at creating one of the most useful tools the discipline has known for keeping its members in touch with each other, they have embarked on this project to make the raw data of the field available to everyone. As part of that effort, a unified ontology is being proposed (Farrar et al. 2002). This draft proposal is just the kind of effort that needs to occur so that the field can make use of the language material on the internet and make progress in its own task of describing languages. It will not be possible for this ontology to satisfy everyone, but the notion of simply having multiple ontologies with translations between them would lead us back to the same kind of regionalism that has given us the current state of affairs in linguistics. It will be much more rewarding to see what evidence guides the arguments over definitions and scope of the terms in the ontology. We can expect this debate to last a considerable time – I would expect at least a decade – but I also expect that there will be greater consensus afterwards than the field has ever known.

There are two aspects to ontology, as Farrar et al. point out in their paper: The first is the definition of each narrowsense grammatical concept and its possible instances, such as DurativeAspect and InstrumentalCase. This will involve laying out the meaning of all the terms in use with a thoroughness seldom attempted. The second is to provide the basis of the analysis of the open-class vocabulary of human languages, which clearly differ in the range of meanings that can be associated with them. Both goals have to be pursued at the same time because the distinction between features for open and closed classes is not clear for any one particular language and can be controversial when unrelated languages are compared. As Farrar et al. (2002: 18) point out, "the features involved in the analysis of the closed class of 'handling verbs' in Athapascan languages are very much like those needed to describe openclass vocabulary items in other languages." Even knowing how the grammatical and lexical features can distribute themselves across languages is something that is not realistically possible now, but will become so as the internet data come into being.

2.5 Summary of effects of internet access to data

The computerization of language data, then, will soon bring us to an unprecedented state of facility for analyzing language. The availability of the raw data

will force a higher standard of descriptive adequacy than we have known for spoken material (as opposed to written). The kinds of searches that can be done will be extraordinarily more powerful than the ones that are available without computerization. The use of sound files for exchanging data will allow better phonological descriptions, ones that take the actual realization of phonemes into account, and will further change the way we do syntax, by making it easier to evaluate and incorporate intonation into the description. And we will be forced to attempt a unified ontology, in order to make the data as useful as possible. This last concern leads us directly to the changes in linguistics that will be more specifically related to endangered languages.

3. Results from endangered languages

The greatest diversity in language features occurs in the languages spoken by the smallest percentage of the world's population. Every language is full of wonders and infinite nuances. Even English, a language of which I am rather fond, has not yielded up all its secrets to the hundreds of linguists who have been poring over it for the past 50 years and beyond. Yet the most extreme cases of language differentiation are surely to be found in the highly divergent language families, many of which are faced with extinction as an entire family, not just a member language or two. Most of the Australian languages, for example, are moribund. Many of them, however, are undergoing revitalization efforts, and some of these may succeed. But entire lines of native L1 learning of these languages has ceased, making it virtually impossible to do new linguistic research on some of the most unusual language families.

This pattern is repeating itself throughout the world, and linguists have been strangely passive about it. Despite the explicit statement I quoted earlier from Bloomfield, the field as a whole developed in such a way that there was little attention paid to the issue of language disappearance, or, more accurately, languages going to sleep. The Australian linguistics community was energized in the early 1980s to begin a systematic recording of all the Australian lan guages, and this project succeeded with a thoroughness that has not been replicated elsewhere. The issue of endangered languages was on the plenary-session level of the agenda of the XV International Congress of Linguists, 1992 (Crochetière et al. 1993). In the U.S., the publication of a group of articles in the journal *Language* (Hale et al. 1992) was the first sign of a collective awareness of the issue within the discipline of linguistics. Soon after that, various groups arose specifically devoted to the preservation of endan-

gered languages. The Linguistic Society of America created the Committee on Endangered Languages and their Preservation (CELP). The National Science Foundation in the U.S. began specifically soliciting projects on endangered languages. In Japan, Osahito Miyaoka obtained funding from the Grant-in-Aide for Scientific Research on Priority Areas of the Japanese Education Ministry (Monbu-kagaku-sho), and for three years, he headed a large scale project with over 150 linguists to document endangered languages of the Pacific Rim (2000–2002) <www.Elpr.bun.Kyoto-u.ac.jp>. Joanna Hess created the Institute for the Preservation of the Original Languages of the Americas, or IPOLA, which has since changed its name to the Indigenous Languages Institute (ILI). The group Terralingua was formed to highlight the interaction between ecosystem and language. In England, the Foundation for Endangered Languages was started by Nick Ostler, and in Germany, Hans-Jürgen Sasse started the Gesellschaft für bedrohte Sprachen. And I started the Endangered Language Fund in the U.S. These groups have similar aims with different emphases, but they are all dedicated to making the most of our linguistic heritage.

Since those early additions to the preservation effort, we have had two substantial infusions of research money. The first was an initiative by the Volkswagen Stiftung in Germany, which promoted the comprehensive description of endangered languages by interdisciplinary teams of researchers. That project is ongoing, and has contributed a great deal to the work on the topic, both in terms of promoting work in the field and in the use of computerization to share the results. Even more recently, the Lisbet Rausing Charitable Fund has initiated a program of grants to support the documentation of endangered languages, and has appointed the School of Oriental & African Studies, London University [SOAS] to administer the scheme. This well-funded program will support a great many projects throughout the world, and the material will be archived at SOAS indefinitely. Here again, we can expect great things to come from this program. In addition, other foundations are beginning to take notice of the issue. The concern does not arise from primarily linguistic concerns but rather for issues of human rights and social cohesion. Still, I think that linguists can be proud that our efforts are resulting in greater awareness of the impact of language on the welfare of various human populations, especially the smaller groups. UNESCO's Intangible Cultural Heritage Unit is renewing its efforts in supporting endangered language communities, and a group of linguists has been working with UNESCO to formulate guidelines and recommendations on maintaining linguistic diversity. A proposal "Language Vitality and Endangerment" is being prepared for the international meeting in March, 2003, by a group that include Matthias Brenzinger, Arienne Dwyner,

Colette Grinevald, Michael Krauss, Osahito Miyaoka, Osamu Sakiyama, Rieks Smeets, Ofelia Zepeda, and Akira Yamamoto. Yamamoto is attempting to mobilize as many linguists as possible by using internet networks in preparing this document.

The focus of this paper, as I said before, is on the other side of this two-way street, the way in which the field of linguistics will change with the study of endangered languages. Clearly, nothing will happen if the languages are not studied, and the support for their study is a necessary component of this picture. The researchers need to be in place for anything to happen as well. And the results have to be made accessible, as I discussed in the first half of this paper. Now I want to discuss some of the ways that endangered languages are going to contribute that would not be served by unendangered languages:

1. The time-scale at which we can understand linguistic change will be greatly expanded if we have data from the languages and families that are going to sleep.
2. Claims for universals and what is learnable will be more justifiable with the more unusual data.
3. The effects of writing, and the limitations of studying only languages with writing systems, can be determined.
4. The linguistic ontology (coming full circle) will be more complete.

None of these benefits can be had once the languages are gone. And literally hundreds of languages will go to sleep during the professional career of those starting out in this year of 2003.

3.1 Increasing the time depth of comparative work

Comparative data have given us the most solid results in the realm of linguistics. The proposed genetic relationships based on correspondences among cognates have withstood the test of time, and the techniques have hardly been improved upon (Hock & Joseph 1996). To tackle greater time depths, we need to have a larger set of relations to be established among farther flung families. This has so far eluded us, and it may never happen. But the use of material from distantly related families is the only hope we have for dealing with a greater time scale than is currently possible. The data that will be required is primarily lexical, but the semantic changes that can be used to establish cognates need to be delimited better than they currently are in order to avoid overgeneralizations. Again, we may never achieve the time depth that we would like, but the data for even trying is disappearing.

3.2 Tests for the essential character of the language faculty

Second, the nature of the language faculty and how it allows for first language acquisition are hotly contested, and there is much more concrete evidence that is necessary. If learning a language is changing settings of a universal set of parameters (e.g., Culicover 1997), we need to know just how broad a range of settings is possible. If learning is only pattern detection, as proposed by neural network theory (e.g., Elman et al. 1996), we need to see if there are in fact patterns that may not be truly learnable. None of this will critically require studying acquisition of the endangered language, though that would, of course, be useful. (The status of most endangered languages as not being learned from the cradle means that the study of acquisition in those languages is not typically possible.) But just knowing what the end-result grammars are like will help us understand what it is that infants must derive as they acquire their language. These issues are widely debated, but there may be a way of putting the question that could help, one that helps us envision something outside the language faculty.

What would it take for a language to be unlearnable by some portion of the human population? One case might be phonological, based on some peculiarity of anatomy. Let's say that some group developed a flap connected to an air sac in one cheek. This might allow for a quick pop to be released any time during a speech sound, even in the middle of a vowel. Humans with the current anatomy would be unable to produce these sounds, and might not even be able to perceive them correctly. In particular, it might be impossible for them to tell when the pops occurred within the speech stream and thus to decide what order the phonemes occurred in. Some unpublished research that Alvin Liberman and I performed to look at timing in speech suggests that this is plausible. Listeners were unable to report where within a syllable a nonspeech click occurred, even when it was exactly in the center of the syllable. They most often reported it at the beginning or the end of the syllable, no matter what its real location was. The ordering of the speech sounds is easy for us because we have a specialization to deal with these sounds (Liberman & Whalen 2000), not because we are generally good listeners.

So if one subspecies of humans developed both the ability to produce these cheek pops and the ability to locate them relative to the total syllable, it might make it impossible for other groups to fully master that language. Of course, one would assume that many aspects of the language would be in common with existing languages, but these new sounds would probably be common in the language. Languages that invest in unusual consonant types appear to rely

on them heavily (Hombert & Maddieson 1998). For example, English has the relatively rare dental fricatives theta and edh, but they are used in many of the most common words of the language. The clicks of the Khoisan languages are not relegated to an occasional word but rather appear throughout the vocabulary. The extremely rare voiceless lateral affricate (occurring in 5 of the 317 languages in Maddieson 1984) is used heavily in Nahuatl. For example, 2985 of the 8306 entries of Amith's online Nahuatl dictionary contain this segment (Amith 1999). We can expect, then, that once this unusual capability was available, it would be heavily used in the language, and the language would be, in a very real sense, unlearnable by other humans.

It may be that brain evolution (where small changes can have large effects) is even faster than evolution in external anatomy, and thus we might see changes there first. Imagine that a language – let's call it Octavian – developed so that grammaticality required exactly one member of eight sets of morphemes to be uttered each day. Now, this certainly exceeds the definition of grammar as relating only to sentence-level phenomena, but the rules of anaphora and coherence clearly require that extension in any event. So, let's say there are eight classes of morphemes, each with real semantic content, but which are mutually exclusive on any particular day. They would also have to occur in order, requiring a great deal of planning to get it right. While this scheme seems fairly outlandish, I think it is fairly easy to see how it might arise. Imagine that there is an extremely strict religion that requires eight prayers during the day. Each prayer should be selected from some set of prayers, as dictated by season or circumstance. Over the years – or centuries, if need be – it is not hard to imagine that the prayers might be reduced to simple words or even bound morphemes, much in the way that "God be with you" ends up becoming "bye." It is also not hard to see how someone who did not have sufficient long-term memory to keep track of these prayers might be excluded from the gene pool. Even if this religion did not strictly punish people for getting it wrong, it still would tend to mark those individuals who got it wrong as "profane" or even just "unlucky." Again, over the years, such pressures could lead to the exclusion of the memory-challenged from the gene pool.

So, picture the difficulties faced by the "normal" language learner who is thrust into the "Octavian" environment. While the Octavians have a sort of memory buffer with eight slots that makes the day's duty seem natural, the "normal" learner is baffled. She can't understand why people keep asking her at bed time if she doesn't have anything else to say. All of her sentences are grammatical, but she goes to bed with an asterisk hanging over her head, punctuating yet another ungrammatical day. Now in this case, it should be possible

to learn what the eight morphological categories are, and what their (daily) distribution has to be. But would a non-Octavian be able to learn this pattern without explicit instruction? Currently, there seems to be no pattern of language (other than the auxiliary system of reading) that requires explicit instruction. This is one of the great strengths of the language faculty, since every normally developing human is a fluent native speaker, but far fewer are good analyzers and teachers. So it might just be impossible for a non-Octavian to learn this daily structure naturally. This would then be another way in which a language might develop so that it would not be learnable by the general population.

There are no doubt other ways in which languages might evolve so that they could not be learned. It is a topic that has not received much attention. Indeed, even the assertion that all children can learn any language is more an article of faith than an established fact. We have only recently seen something like a natural experiment of this, in the long-distance adoption practices of the modern world. The universality of language does seem to hold up well under this test, but are there subtleties that are missed? The languages that have large enough groups of speakers to support this long-range adoption of infants may simply have lost anything that might be unique. If a genetic change affected a small group and *their* language was not learnable, we would never know that fact if it was their children that were adopted into a majority group. So the natural experiment is incomplete, as are all natural experiments. To the extent that it is possible to extend that experiment, the time is now, while we have over 6,000 languages to work with. The languages that are most vulnerable are exactly those that are the most likely to have diverged significantly.

If there is already a language out there that is not fully learnable by the majority of the human species, we may not have the tools to discover it. An examination of the discussion of slips of the tongue provides a good analogy here. It has often been claimed that slips are segmental and mostly phonotactically coherent (e.g., Fromkin 1971). Yet when the data consist of transcriptions that are done in a nearly exclusively segmental system like the IPA, it is hardly surprising that segments seem to be the proper level of analysis. More recent studies that have looked at muscle acti vation (Mowrey & MacKay 1990) or movements of the articulators (Pouplier & Goldstein 2002) find consistent subsegmental effects that are often inaudible. Only when we look at the signal with the right tool do we see the unusual. We may already be missing features of languages that have evolved beyond the normal ability.

3.3 Effects of literacy

The third area that endangered languages can uniquely inform us on are the effects of literacy. Most of the endangered languages have never had an orthography designed for use by native speakers, and many of those that have had such orthographies devised do not make extensive use of them (e.g., Bielenberg 1999). There are many reasons that such a lack of use can occur. There may not be enough technical support for printing, paper, computers, etc. The community may not feel that there is enough need to invest the time in learning how to use an orthography. Dialect differences may result in "orthography wars" which make the simplest writing a politically sensitive act. And some groups simply feel that writing destroys the spirit of the language. It is the scientific roots of this last feeling that I would like to explore here.

Writing changes a language, and some languages resist writing. These are not self-evident facts, or, at least, I hope they are not self-evident, because it took me a long time to accept them. Writing is parasitic on speech, as we know (e.g., Coe 1992). There is no language community in which writing has replaced speech or sign as the primary means of communication. Children do not acquire reading and writing as their first language. Yet this does not mean that writing must necessarily leave the language system intact when it begins its parasitic existence. (Parasites are typically bad for the host organism, but I do not mean to imply that here; I am very happy to be literate, and I am glad that you are able to read my words as well.) Learning to read changes the way words are stored in the mental lexicon. When English speakers make auditory judgments on rhyming words, they are affected by the spelling even though it is never presented in the experimental setup (Seidenberg & Tanenhaus 1979). Learning to read also affects which part of the brain processes phonologically legal nonwords (Castro-Caldas et al. 1998). Writing makes a language more conservative and makes older forms available on a regular basis, both enriching and complicating the language environment. So, the notion that a language should not be written because writing changes, and often fossilizes, a language is one that has some support. Only when we examine the trade-offs involved in adopting writing as a major part of the language community can we determine whether a literacy campaign should be started.

Some languages are resistant to writing as well, and it could easily be that they would change dramatically if they were written on a regular basis. Most of the time, the sound system is not a major obstacle to literacy. The 95 consonants and 24 monophthongal vowels of !Xu~ (Maddieson 1984) are a challenge, though an apparently successful orthography was devised (Lloyd

1911). Yet some instances are problematic. Tarahumara is legendary for its phonetic variability (Copeland 1994). Literacy has not gained much ground within the native community, but some speakers have become literate in their second language, Spanish. Amazingly, these speakers then lose their variability in their native Tarahumara, even though they are not writing it (Copeland, personal communication). Literacy can have large effects, probably the more so as the language is atypical.

The syntax of a language can apparently change as well. Kalmar (1985) reports a case in which Inuktitut apparently created a new class of subordinate clauses due to increased literacy in the community. Although many of the speakers were bilingual in English, it did not appear that the new construction was simply a borrowing. Rather, tendencies within the language appear to have been heightened by the use of writing and subsequently entered the spoken language.

It is quite possible that languages change in other ways as they make greater use of literacy and incorporating a description of those changes will entail a major change in our conception of language. We are at a convergence of trends that would allow us to gain a better understanding of just what the increase in literacy means for the use of the language as a whole. Some of the currently endangered languages that will survive will probably do so by making greater use of literacy, which is a seemingly indispensable tool in dealing with the ubiquitous presence of the global economy. Linguists today will therefore be able to study this process in a way that may never be available again. It would be especially useful if several different languages entering into similar literacy programs were studied at once, allowing for better generalizations across the necessarily divergent starting conditions. A study of what it is that is difficult about writing a language like Tarahumara could allow for predictions to be made for specific languages, which would be testable within a generation of readers. It is so seldom that linguists are able to make predictions about language change that this would be an extremely appealing opportunity for research. And, again, to make a bit of a statement about the other side of this two-way street, it would allow the linguist to make recommendations for the literacy program.

3.4 Diverse inputs to the ontology

The fourth feature that endangered languages offer linguistics is their typological diversity and its effect on the ontology. We can take any three languages and divide up their grammatical categories so that a coherent story can be told. If

the three languages all happen to come from a literate tradition, we might expect that their grammars would already be molded in some similar ways. It is only when we add in the features of the typologically distant and unusual languages that our theories really get a workout. The task of collecting the relevant data has been so large that only limited, though impressive, progress has been made on these issues. If we are ever to be able to listen to what these endangered languages have to tell us, now is the time. They are not going to be talking for that much longer. And they will only be sleeping rather than vanished if we use the tools of our trade to help communities preserve their languages.

The task is large, and it would benefit from efficiency gains in the way we do linguistics, but there has been very little in the way of such gains so far. The use of the internet will be one clear case. The ultimate tool will be the one that will take a recording of a text, do automatic speech recognition on it, return a transcription, perform a morphological mark-up, and then give a literal and free translation. I regret to say that this ideal will not come about before most of the languages spoken today have fallen silent. Anyone who simply records texts with no accompanying translation, in the hopes that they will be automatically translated by a machine of the future, may be passing that hope on to her grandchildren's grandchildren. Texts must be translated now, as they are collected, and the more mark-up that can be accomplished today, with the assistance of native speakers, the more useful that text will be for all time and for heritage learners as well as linguists.

The one real gain in the speed of analysis has been the training of native speakers as linguists. This has been successfully done in all parts of the world, and it results in much more progress for both sides of the equation than would be possible without that training (Hardman 1985). Other groups, such as the American Indian Language Development Institute and the Oklahoma Native Language Association, are also training native linguists. We can also hope that as our field begins its current revolution, the tools will be easier to teach (being based on more accurate and extensive data) and easier to apply in understandable and consistent ways.

4. Conclusion

In short, we are poised to see a revolution caused by an unprecedented level of access to the raw materials of our discipline, using tools that have only recently become available. The very act of creating the tools to access these data is part of the revolution itself, as epitomized by the search for a unified linguistic on-

tology. The most important part of this revolution is that the elements of the linguistic periodic table (to continue the analogy to chemistry) will become clearer, and the standards of evidence for proposing or discarding these elements will become clearer as well. The minority languages of the world have a tremendous amount to offer linguistics in this regard. Again, let me emphasize that I have only been talking about one side of a two-way street in the paper. Linguistics has much to offer the native communities as well, and linguists are much more aware of the ethical ways of obtaining, analyzing and sharing their results than they have ever been. There have been many cases in which later generations have felt the loss of their language and want to waken it from its sleeping state, and they have benefited from the linguist's efforts at preserving it. We can be sure that this will happen repeatedly in the coming decades.

We have reached a stage in the study of language at which it is no longer ethical for a linguist not to consider working on an endangered language. This is not to say that all linguists must work on an endangered language: There are many valid reasons why a linguist might decide that a nonendangered language is the most appropriate one to study. But not to ask the question is insupportable. However, it is not just a moral issue: The vanguard of the revolution will be those who study endangered languages. It is always a heady decision to join a revolution, but the rewards of a successful one are great. And we are witnessing the beginning of a successful revolution in linguistics.

Acknowledgements

I would like to thank Christian Lehmann, along with the Comité International Permanent des Linguistes, Ferenc Kiefer, Chair of the Scientific Committee, and Piet van Sterkenburg, Secretery General of CIPL, for inviting me to give this presentation. The writing of this paper was supported by NIH grants HD-01994, DC-00493, DC-03102, and DC-02717 to Haskins Laboratories. I thank Steven Bird, Scott Farrar, Carol A. Fowler, Peter Ladefoged, Christian Lehmann, Simon Levy, Joyce McDonough, Nick Ostler, Hans-Jürgen Sasse, Michael Studdert-Kennedy, Tasaku Tsunoda, Mary Willie, and Akira Yamamoto for helpful comments; they are not responsible for any lapses that remain.

References

Aissen, J. (1997). "On the syntax of obviation." *Language, 73*, 705–750. http://www.yale.edu/nahuatl/index.html

Bielenberg, B. (1999). "Indigenous language codification: Cultural effects." In Jon Reyhner, Gina Cantoni, Robert N. St. Clair, & Evangeline Parsons Yazzie (Eds.), *Revitalizing indigenous languages* (pp. 103–112). Flagstaff, AZ: Northern Arizona University.

Bird, S. (1999). "Multidimensional exploration of online linguistic field data." In P. Tamanji, M. Hirotani, & N. Hall (Eds.), *Proceedings of the Northeast Linguistics Society* (pp. 33–50). Amherst: University of Massachusetts.

Bloomfield, L. (1925). "Why a linguistic society?" *Language, 1*, 1–5.

Bloomfield, L. (1962). *The Menomini Language*. New Haven: Yale University Press.

Castro-Caldas, A., Petersson, K. M., Reis, A., Stone-Elander, S., & Ingvar, M. (1998). "The illiterate brain: Learning to read and write during childhood influences the functional organization of the adult brain." *Brain, 121*, 1053–1063.

Coe, M. D. (1992). *Breaking the Maya Code*. London: Thames and Hudson.

Copeland, J. (1994). "Variation and constancy of patterning in language and culture: The case of Tarahumara." In Valerie Makkai (Ed.), *The twentieth LACUS forum* (pp. 5–30). Chapel Hill, NC: Linguistic Association of Canada and the United States.

Crochetière, A., Boulanger, J.-C., & Quellon, C. (Eds.). (1993). *Endangered languages: Proceedings of the XVth International Congress of Linguists*. Quebec: Presses de l'Université Laval.

Crystal, D. (2000). *Language Death*. Cambridge: Cambdridge University Press.

Culicover, P. (1997). *Principles and Parameters: An Introduction to Syntactic Theory*. Oxford: Oxford University Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Farrar, S., Lewis, W. D., & Langendoen, D. T. (2002). "An ontology for linguistic annotation." In *Semantic Web Meets Language Resources: Papers from the AAAI Workshop, Technical Report WS-02-16* (pp. 11–19). AAAI Press.

Frantz, D. (1966). "Person indexing in Blackfoot." *International Journal of American Linguistics, 32*, 50–58.

Fromkin, V. A. (1971). "The non-anomalous nature of anomalous utterances." *Language, 47*, 27–52.

Grenoble, L. A. & Whaley, L. J. (Eds.). (1998). *Endangered Languages: Language Loss and Community Response*. Cambridge: Cambridge University Press.

Hale, K., Krauss, M., Watahomigie, L., Yamamoto, A., Craig, C., Masayesva Jeanne, L., & England, N. (1992). "Endangered languages." *Language, 68*, 1–42.

Hardman, M. J. (1985). Linguistics and native control of language policy. Ms., *Presented at the 15th anniversary of the Aymara Langbauge Materials Program at the University of Florida*. Gainesville, FL.

Harris, R. A. (1993). *The Linguistics Wars*. New York: Oxford University Press.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). "The faculty of language: What is it, who has it, and how did it evolve?" *Science, 298*, 1569–1579.

Hinton, L. (2001). "Sleeping languages: Can they be awakened?" In Leanne Hinton & Ken Hale (Eds.), *The Green Book of Language Revitalization in Practice* (pp. 413–417). London: Academic.

Hinton, L. & Hale, K. (Eds.). (2001). *The Green Book of Language Revitalization in Practice.* London: Academic.

Hock, H.-H. & Joseph, B. D. (1996). *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics.* Berlin: Mouton de Gruyter.

Hockett, C. (1966). "What Algonquian is really like." *International Journal of American Linguistics, 32,* 59–73.

Hombert, J.-M. & Maddieson, I. (1998). "A linguistic approach to automatic language recognition." *UCLA Working Papers in Phonetics, 96,* 119–125.

Kalmar, I. (1985). "Are there really no primitive languages?" In David R. Olson, Nancy Torrance, & Angela Hoildyard (Eds.), *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing* (pp. 148–166). Cambridge: Cambridge University Press.

Lederer, R. (1989). *Anguished English: An Anthology of Accidental Assaults Upon our Language.* London: Robson.

Liberman, A. M. & Whalen, D. H. (2000). "On the relation of speech to language." *Trends in Cognitive Sciences, 4,* 187–196.

Lloyd, L. C. (1911). "A few !Kun texts." In W. H. I. Bleek & L. C. Lloyd (Eds.), *Specimens of Bushman folklore* (pp. 404–433). London: George Allen & Unwin.

Maddieson, I. (1984). *Patterns of Sounds.* New York: Cambridge University Press.

Maffi, L. (2000). "Language preservation vs. language maintenance and revitalization: Assessing concepts, approaches, and implications for the language sciences." *International Journal of the Sociology of Language, 142,* 175–190. http://ucjeps.herb.berkeley.edu/Maffi-book.html

Maffi, L. (Ed.). (2001). *On Biocultural Diversity: Linking Language, Knowledge and the Environment.* Washington, DC: Smithsonian Institution Press (see also Amith, J. (1999). Analytic dictionary of Ameyaltepec Nahuatl).

Martinet, A. (1960). *Éléments de Linguistique Générale.* Paris: Armand Colin.

Mowrey, R. A. & MacKay, I. R. A. (1990). "Phonological primitives: Electromyographic speech error evidence." *Journal of the Acoustical Society of America, 88,* 1299–1312.

Pouplier, M. & Goldstein, L. M. (2002). "Asymmetries in speech errors: Production, perception and the question of underspecification." In T. A. Hall, Bernd Pompino-Marschall, & Marzena Rochon (Eds.), *Papers on Phonetics and Phonology: The Articulation, Acoustics and Perception of Vonsonants* (pp. 73–82). Berlin: ZAS.

Sasse, H.-J. (2002). "Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state?" *Linguistic Typology, 6,* 199–271.

Seidenberg, M. S. & Tanenhaus, M. K. (1979). "Orthographic effects on rhyme monitoring." *Journal of Experimental Psychology: Human Learning and Memory, 5,* 546–554.

Whalen, D. H. (2001). "How can we ethically put language material on the web?" *The Endangered Language Fund Newsletter, 5* (1), 1–4.

Linguistics Today - Facing a Greater Challenge

Edited by

Piet van Sterkenburg
Institute for Dutch Lexicology, Leiden

John Benjamins Publishing Company
Amsterdam/Philadelphia

Library of Congress Cataloging-in-Publication Data

Linguistics Today : facing a greater challenge / edited by Piet van Sterkenburg.
p. cm.
Includes bibliographical references and indexes.
1. Linguistics. I. Sterkenburg, P. G. J.

P121. L5713 2004
410-dc22 2004050174
ISBN 90 272 3223 7 (Eur.) / 1 58811 537 2 (US) (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA