Abstract

If global social integration continues, it appears likely that languages will keep dying until the world is unilingual. The implementation of cultural, economic, political, and technological strategies of panlingual globalization might prevent that outcome. One such strategy, aiming at panlingual transparency, is currently being investigated in a research project at the University of Washington's Turing Center. The researchers are developing a panlingual lexical database, inference algorithms, and practical applications, to discover whether panlingual lexical translation can help breathe new life into thousands of doomed languages by making them serve their speakers' information and communication needs. Initial results of experiments support the belief that the world's dispersed lexical resources can be joined into a panlingual translation engine, which people can use effectively in some representative Web-search and interpersonal-messaging tasks. Efforts to make the engine accessible for worldwide research, development, and use are beginning.

Panlingual Globalization

Jonathan Pool

Predicting Unilingual Globalization

The complex relationship between globalization and linguistic diversity (Mufwene 2004) makes it difficult to predict the changes in the distribution of languages that will accompany future advances in world social integration. Figure 1 shows one highly simplified idea of their relationship. Here progress in information and communication technology (ICT) is modeled as promoting global interactivity among communities, and this in turn encourages shifts from low-density (smaller and less resource-endowed) second and native languages to high-density ones. This causal relationship would make one expect a decline in linguistic diversity as globalization proceeds. However, the same technological progress facilitates the development of tools and resources usable for the maintenance and cultivation of low-density languages and the creation of viable communities out of linguistic diasporas. Such progress could allow linguistic diversity and globalization to thrive together.
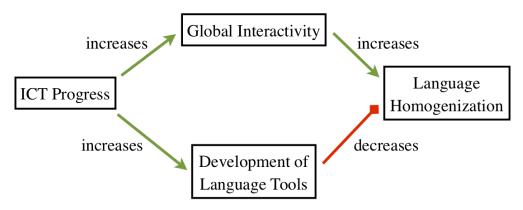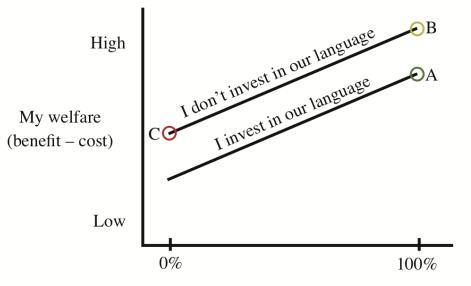
```
         increases    Global Interactivity    increases
      ┌──────────────→ ┌──────────────────┐ ──────────────→
      │                └──────────────────┘                ┌──────────────────┐
 ┌──────────────┐                                          │     Language     │
 │ ICT Progress │                                          │  Homogenization  │
 └──────────────┘                                          └──────────────────┘
      │                ┌──────────────────┐
      │  increases     │  Development of  │    decreases
      └──────────────→ │  Language Tools  │ ──────────────
                       └──────────────────┘
```

Figure 1. Globalization and Unilingualization.

If globalization can both promote and diminish linguistic diversity as shown in Figure 1, the net impact of globalization may depend on human motivations. The more the world's population wants to participate in linguistic diversity and the more the native speakers of low-density languages want to maintain and transmit them, the more they will exploit ICT for these purposes, and thus the more directly globalization and

linguistic diversity will co-vary.

Most of the evidence seems to predict an inverse relationship, because linguistic diversity, maintenance, and revitalization are not generally popular ideals. Low-density languages throughout the world have been dying, only rarely showing resistance (UNESCO 2003: 2–4). Typically, parents do not demand that these languages be transmitted to their children; children do not insist on learning them; and schools do not require pupils to learn them. Often, speakers of low-density languages even try to prevent their children from learning and using them, in part because they are under the influence of denigrating opinions held by outsiders (Eidheim 1969; Harrison 2007). The world's population, as a whole, treats low-density languages as inferior, or at best superfluous (Crystal 2000: 27). People decide, given this opinion, that assimilation within and across generations to high-density languages confers net benefits on those who assimilate, if the cost of assimilation is not excessive. Globalization decreases that cost by creating opportunities for immersive learning of high-density languages. These combined forces have led to predictions that something between half and 90% of the world's living languages will die within the next century (Woodbury 2006; UNESCO 2003: 2). Weaker forms of these forces appear to be shrinking the use of medium-density languages in science, diplomacy, business, and other domains (Phillipson 2008).

Figure 2. Low-Density Language Dilemma.

Even if linguistic diversity became much more popular, this change might not suffice to produce a positive globalization–diversity relationship. Suppose that, in general, any benefits conferred by linguistic diversity were dispersed but all its costs were imposed on those who maintain low-density languages. In other words, suppose that the choice whether to learn, use, document, and enrich low-density languages took the form of a collective action dilemma, each native speaker of such a language finding himself/herself in a situation modeled by Figure 2 (see De Swaan 2004: 579). In this dilemma, if everybody cultivates the language everybody is at A, and if nobody does so everybody is at C. Everybody prefers A to C. But any individual at A can reach B, thus enjoying increased benefits, by defecting (not cultivating). If all individuals yielded to that incentive, the outcome would change and everybody would be at C. The language would probably atrophy and die.
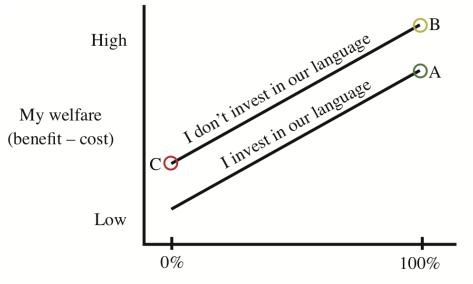
Strategies for Panlingual Globalization

Those who reluctantly predict linguistic homogenization accompanying globalization need not simply despair; they can try to render their prediction false. Consider the following examples of action strategies.

Strategy 1: Marketing Multilingualism. Persuade the world's population that the existence of about 7,000 languages (Gordon 2005) is a boon rather than a curse. This is the strategy attempted by Nettle and Romaine (2000), Crystal (2000), Abley (2003), and Harrison (2007). When languages die, they argue, the world loses (1) irreplaceable knowledge of history, medicine, nature, and productive methods encoded in languages'

lexicons; (2) evidence for the scientific understanding of language and the human mind; (3) diverse ideas arising from languages' differing systems of knowledge representation; and (4) the respect, tolerance, sophistication, and enjoyment that develop (or could develop) from people learning to live in a multilingual world. They further argue that cultural and biological diversity and diverse identities, all of which are already widely appreciated, depend on linguistic diversity, which should therefore be valued for its effects even by those who do not value it intrinsically.

This strategy, if effective, would make the world want linguistic diversity, but that want would not by itself stop the erosion of linguistic diversity. An increased popular appreciation of linguistic diversity might merely make the slopes in Figure 2 steeper, as in Figure 3. In this case the predicted (equilibrium) outcome would be the same, and overcoming the dilemma would require additional strategies.

Figure 3. Low-Density Language Dilemma with Diversity Popular.

Strategy 2: Ecolinguistic Compensation. Design mechanisms to internalize the benefits of low-density language cultivation. This strategy would give financial support to those who keep their native languages alive and vibrant. The world could absorb the costs of the analysis, documentation, instruction, and other activities that the cultivators require (UNESCO 2003). Beyond that, the world could treat active native speakers of low-density languages as service providers and pay them compensation.

Consider a numerical example. Suppose that keeping the 5,000 lowest-density languages alive and vigorous costs $5,000,000,000 per year ($1,000,000 per language per year) and yields benefits (knowledge, identity, tolerance, and so on) worth $30,000,000,000 per year (1⁄20 of 1 percent of the gross world product). If the native speakers of those languages total 1,000 persons each, or 5,000,000 altogether (a plausible estimate, given that only about 400 languages have 10,000 speakers or more), then each native speaker incurs a cost, on average, of $1,000 per year. Let us assume that they have no special affection for their native language, so they share the benefits of their cultivation equally with all the others in the world. If so, the annual benefit enjoyed by each native speaker is $5 ($30,000,000,000, split among the 6,000,000,000 persons in the world).

In this example, without a subsidy native speakers who cultivated a low-density language would incur a cost of $1,000 for a benefit of $5 annually. An ecolinguistic compensation policy could pay the maintainers of a low-density language $2,000 per year each. This would give them 200 percent returns on their investments, while still leaving the rest of the world with a $20,000,000,000 annual net benefit ($30,000,000,000 in gross benefit minus $10,000,000,000 in compensation costs).

Compensation mechanisms have been analyzed as a means of making dominant languages more equitable for those who do not speak them natively (Van Parijs 2007) and of making official-language policies fair and efficient (Pool 1991; Ammon 2006: 333–336). A close parallel is that of ecological compensation mechanisms (also known as payments for environmental services, or markets in biodiversity services); these have been in use for about thirty years (Ferraro andi Kiss 2002; Jenkins et al. 2004).

Strategy 3: Linguistic Subsidiarity. Reorganize social life to make linguistic communities more self-governing and socioeconomically autonomous. This strategy would aim to make the world more like a community of language communities than like a community of nation-states, territories, religions, ideologies, or other subpopulations. A self-governing and internally cohesive low-density language community could make its language official and treat it as the main medium of education, commerce, publication, and other social interaction, more easily than is possible where the language is merely that of a minority. With the progress of telecommunications, non-contiguous communities, such as those formed by linguistic diasporas, become more feasible. The strategy would not only make jurisdictional and transactional boundaries more coincident with language boundaries, but also, as proposed by Bastardas (2002), transfer authority from world bodies to single-language local units of government as much as is practical (the subsidiarity principle), thereby making the languages of those units useful and used. According to Mufwene (2002), utilization, particularly in a person's work, is the critically necessary condition for the survival of a low-density language.

Strategy 4: Panlingual Transparency. Create language processing systems that automatically translate utterances among all the languages of the world. This strategy would attempt to produce a real-world counterpart to the fictional "Babel fish" of Adams (1979: 51–52). Such systems would allow anybody who knows any language to understand thoughts and emotions expressed in any other language. In this situation, the incentives for assimilation to high-density languages would be diminished, with the amount of diminution depending on the quality of the translation.

Research and development in machine translation have been active for about half a century (Hutchins 2006; Trujillo 1999), the goal almost always being to translate between particular pairs or small sets of (generally high-density) languages. A few systems under current development apply to larger sets of languages, but never to more than about 50 (e.g., http://translate.google.com; http://www.langtolang.com).

Attempts to realize panlingual—or even massively multilingual—translation have mostly involved human effort rather than automatic processing; these projects have mainly focused on particular bodies of text, such as the Universal Declaration of Human Rights (UDHR 2008), and the user interfaces of particular computer application programs, such as search engines (for instance http://www.google.com/support /bin/static.py?page=searchguides.html&ctx=preferences&hl=en#searchlang). However, some approaches to language modeling, including machine-translation interlinguas (Schubert 1992; Dorr et al. 2006) and typological grammar engineering (Bender and Flickinger 2005), might make automatic translation efficiently extensible to any number of languages.

As these four strategies illustrate, panlingual globalization might be pursued in radically diverse ways (see Fettes 2003; Tonkin 2003). At their simplest, Strategy 1 is cultural, Strategy 2 is economic, Strategy 3 is political, and Strategy 4 is technological. It is plausible that the most effective approach to panlingual globalization would combine these and other strategies, rather than relying on only one.

Engineering Panlingual Globalization

Any strategy for panlingual globalization is likely to arouse doubts because it aims at an outcome which was never experienced and is far from current reality. For example:

1. How could the world be persuaded to value linguistic diversity highly?
2. If ecolinguistic compensation were paid, how could one know who is eligible for the payments and how much to pay each of them?
3. Aren't there far too many entrenched interests aligned with existing jurisdictional boundaries to make linguistic subsidiarity achievable?
4. Don't the still laughable automatic translations between high-density languages, after half a century of effort, show that panlingual translation is simply too difficult?
5. More fundamentally, might efforts to preserve low-density languages inadvertently devalue medium-density ones and thereby hasten global unilingualism (De Swaan 2004)?

To evaluate these doubts, one can attempt to implement each strategy. This brings us from the stage of envisioning panlingual globalization to the stage of engineering it. The following discussion will focus on an actual attempt to begin engineering panlingual transparency (Strategy 4).

In late 2006, the University of Washington's Turing Center (http://turing.cs.washington.edu), with the support and collaboration of Utilika Foundation (http://utilika.org), began investigating the possibility of translation among thousands of languages. Even though, as mentioned above, existing automatic translation

systems are limited to about 1 percent of the world's languages, they have produced results far inferior to expert human translations. As one example, consider the translations of an English sentence into French produced by nine systems currently offered to the public, shown in table 1. Ambiguities like "speaker" and "go on", which human translators easily resolve, often defeat machines. (French "haut-parleurs" refers to amplifying devices. French "a continué" and "s'est passé" can be translated "went on", but this sense is not applicable here. If automatic translation is difficult for the most richly endowed languages, there is reason to be pessimistic about automatic translation from every language into every other language.

Table 1. Automatic Translations from English into French.

Role	Text
Source	Both speakers stopped talking after the warning light went on
Target, PITS (http://translation2.paralink.com)	Les deux speakers ont arrêté de parler après que la lumière d'avertissement a continué
Target, SYSTRANet (http://www.systranet.com)	Les deux haut-parleurs ont cessé de parler après que le voyant d'alarme se soit allumé
Target, Babylon Online Translator (http://translation.babylon.com)	Les deux orateurs ont cessé de parler après le voyant s'est passé
Target, Live Search Translator (http://microsofttranslator.com)	Les deux orateurs cessé de parler après que le voyant d'avertissement a
Target, Google Translate (http://translate.google.com)	Les deux intervenants ont cessé de parler après le voyant d'alerte s'est passé
Target, PROMT Translator (http://www.online-translator.com)	Les deux orateurs ont arrêté de parler après que la lumière d'avertissement a continué
Target, SDL FreeTranslation.com (http://ets.freetranslation.com)	Les deux orateurs ont arrêté de parler après que la lumière d'avertissement a continué
Target, Reverso Translation (http://www.reverso.net)	Les deux orateurs(locuteurs) ont arrêté de parler après que le témoin lumineux a continué
Target, InterTran (http://www.tranexp.com:2000/Translate/result.shtml)	Tous les deux interlocuteurs arrêtions parler à la suite les voyant lumineux êtes allé one

After investigating some alternatives, the Turing Center researchers concluded that they could design a system to perform one type of translation more or less panlingually: lexical translation. The system would translate lexemes, the elements of the lexicons (vocabularies) of languages. For example, the system would not translate "Both speakers stopped talking after the warning light went on". Instead it would translate the lexemes "both", "speaker", "stop", "talk", "after", "the", "warn", "light", "go", and "on". It might also translate "warning", "warning light", and "go on", since they, too, may be considered lexemes (they may appear as entries in dictionaries).

This project of panlingual lexical translation ("PanLex") was massively multilingual from the beginning and is rapidly extensible to cover all languages (being limited only by the available data). In compensation, PanLex translates lexemes and makes no attempt to translate sentences, paragraphs, or longer discourses. We might describe it as initially wide but shallow; most translation systems, by contrast, begin deep but narrow. Other systems may be asked, "You don't cover my language, so what good can you do for me?"; PanLex may be asked, "You cover my language, but you translate only lexemes, so what good does that do for me?"

The hypothesis underlying PanLex was that lexical translation is more useful than one might imagine. Some utterance types often consist merely of sequences of lexemes. Web search queries, library-style subject headings, entries in book indices, user-interface labels ("copy", "undo", etc.), social tags on the Web, list entries (places, events, hobbies, interests, and the like), weather-forecast summaries, telegrams, SMS text messages, baby talk, and foreigner talk are among them. Moreover, utterances that generally contain morphology and syntax may be converted to sequences of lexemes, and the sequence and context may make them fully or partly intelligible. Grammatically conveyed information, such as time, number, illocutionary force, or evidentiality, may be expressed with lexemes (such as "yesterday", "many", "question", or "allegedly"), and, if not so expressed, may still be successfully inferred. Even in situations where purely lexical translation is insufficient, it may be easily and inexpensively supplemented; this would result in a family of equivalent controlled languages (Pool 2006) with minimalistic syntax, which would avoid the structural ambiguities of natural languages. For example, communicators might supplement "man, bite, dog" with annotations to specify which of the verb's arguments is the agent, and whether the statement is an assertion, a question, or a recommendation. The idea that simple annotation techniques may have great

expressive power is akin to one of the assumptions of the Semantic Web Initiative (Berners-Lee et al. 2001): that human communication references massive numbers of things but only a few relationships among those things.

PanLex draws on various lexical resources, including dictionaries, wiktionaries, glossaries, lexicons, word lists, terminologies, thesauri, wordnets, ontologies, vocabulary databases, named-entity resources, and standards. Despite their different names and formats, they all assert facts of the type "Lexemes A, B, C, ... , and N share at least one meaning common to them all". The fact that they share a meaning makes them synonyms if they belong to the same language, or translations if they belong to different languages.

There are thousands of these resources in existence, and they report the equivalences of millions of lexemes in thousands of languages. One of the first resources usually bestowed on any low-density language is a dictionary or word list. Such a resource usually translates between that language and some higher-density language, such as English, French, Spanish, Russian, German, Hindi, or Tok Pisin. However, any arbitrary pair or larger set of languages might be covered. For example, there are resources linking Greek with Catalan, Nepali with Esperanto, and Turkish with Azerbaijani. About 300 multilingual resources are being developed in the Wiktionary project (Wikimedia Foundation 2008); each wiktionary has a single source language and translates lexemes into an unlimited set of other languages. There are also specialized resources, sometimes organized as thesauri with taxonomies of meanings expressed in multiple languages; one example is the Food and Agriculture Organization's thesaurus (FAO 2008), which expresses about 28,000 meanings related to agriculture and nutrition in Arabic, Czech, Mandarin, German, English, French, Hindi, Hungarian, Italian, Japanese, Lao, Western Farsi, Polish, Portuguese, Slovak, Spanish, and Thai. Finally, there are monolingual resources (thesauri and wordnets) that identify synonyms.

PanLex defines concepts pragmatically. When a resource asserts that some lexemes share a meaning, PanLex assigns a new identifier to that meaning, leaving for later the question whether it is the same meaning as any meaning from any other resource. The simplest bilingual word lists, such as the one shown in Figure 4, give no information about a lexeme except its lemma (its dictionary or citation form). PanLex accordingly treats a lemma in a language as a lexeme. While some other systems might analyze English "tear" (eye water) as one lexeme and "tear" (rip) as another, PanLex treats "tear" as a single lexeme. More complex resources, like that shown in Figure 5, provide additional facts about lexemes and meanings. PanLex recognizes four fact types that often appear in complex lexical resources: definition, domain, meaning identifier, and word class. It also recognizes a generic fact type, consisting of an arbitrary attribute–value pair. This can be used for otherwise unrecognized facts, such as etymology, argument frame, register, and usage.

Figure 4. Simple Lexical Resource.

پاتا pātā

پاتا *pātā*, s.f. (6th) The funeral service (from A فاتحه) (E.) Sing. and Pl.; (W.) Pl. پاتاوي *pātāwī*. پاتا کول *pātā kawul*, or پاتا ویل *pātā wa-yal*, verb trans. To offer up prayers for the dead, to perform the funeral service, to make an exordium.

Figure 5. Complex Lexical Resource.

PanLex recognizes a range of language varieties. Most are ordinary natural languages, such as Burmese and Zulu, but the system can accommodate ethnic dialects, controlled natural languages (Pool 2006), artificial languages (Blanke 1989; Libert 2000 and 2003), and the controlled vocabularies embodied in standards. For example, the ISO 639 standard (SIL 2008) is treated as a language variety in PanLex. This standard identifies nearly 8,000 three-letter codes to represent the human languages of the world; each code is a lexeme in the ISO 639 language variety.
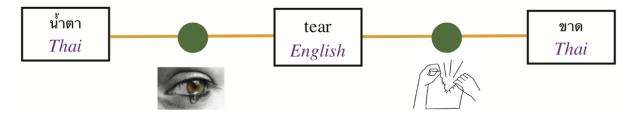

Figure 6. Graphical Interpretation of Denotations.

Table 2. Tabular Interpretation of Denotations.

Meaning	Language	Lexeme
1	English	tear
1	Thai	น้ำตา
2	English	tear
2	Thai	ขาด

Logically, the main facts recorded by PanLex are assignments of meanings to lexemes. These facts, called "denotations" in the PanLex terminology, take the form "authority A asserts that lexeme L has meaning M." From two or more denotations, one can derive assertions about translations and synonyms. If some authority says that lexeme A has meaning X and also says that lexeme B has meaning X, then that authority considers A and B to be translations or synonyms. The entire collection of the denotations can be interpreted geometrically or tabularly. Geometrically, it has the logical form of an undirected graph, as in Figure 6. The graph contains nodes (points) of two types: lexemes and meanings. Edges (lines) represent denotations; each edge connects one lexeme node with one meaning node. If a resource asserts a fact about translations or synonyms, the fact is represented as a single meaning node connected to two or more lexeme nodes. Tabularly, the collection of denotations can be viewed as a three-column table, as in table 2, with each row representing a denotation. An asserted translation or synonymy is represented as two or more rows with the same meaning and distinct lexemes. The denotations are actually stored in a relational database, so that users can efficiently use the system as a translation engine.

Prototypes, Experiments, and Results

In the effort to make PanLex a useful system, its developers have faced three principal challenges.

The **first challenge** has been to collect enough lexical facts from enough language varieties to make PanLex realistically large. About 600 lexical resources have been consulted to date. Although these resources are in machine-readable form, most were created for human readers and rely on the readers' knowledge and intuitions. For example, dictionaries commonly use symbols such as "~" to indicate that a part or all of a headword is to be repeated, but the repeated item may vary irregularly. Translations into phrases containing commas, such as "there, there", are often intermixed in the same resource with translations into multiple synonymous expressions, such as "often, frequently", and translations into disjunctions with shared constituents, such as "soccer, football field". Resources are often constructed over many years, and formats change while the work is in progress. Multilingual resources are often collaborations among persons or teams that follow different conventions of punctuation, capitalization, and orthography. Moreover, a worldwide conversion of character encoding from multiple conflicting systems to a single coherent standard, Unicode (Unicode 2007), has been in progress since 1991, but many digitized resources remain encoded under pre-Unicode standards, some of them poorly defined. Finally, even resources that are consistently organized and well encoded exhibit incompatibilities, such as in diacritical marks and in other aspects of spelling. Automatically combining facts attested by multiple resources requires that, if two facts refer to the same lexeme, the lexeme be identifiable as the same. All of these problems require that extensive normalization be performed on data contributed by resources.

Notwithstanding these obstacles, as of April 2009 (about 2.5 years after the project was launched), the database contains about 27,400,000 denotations. They assign, in total, about 10,100,000 meanings to about 12,300,000 lexemes in about 1,300 language varieties. On the basis of these facts, it is possible to perform about 204,500,000 different translations (102,200,000 pairs of lexemes, each translatable in both directions). Here "translations" include intralingual translations ("Lexeme B is a synonym of lexeme A"), which constitute about 5 percent of the total.

The **second challenge** has been to fill gaps in the data with artificial intelligence. The data provide only a small fraction of the translations that users might want, even among the lexemes already in the database. To get translations from any lexeme into any language variety, users require not only attested facts but also automated inference from those facts. Consider the case in which somebody wants to translate the Icelandic word "hnappur" into Arabic (Figure 7). The database currently assigns three meanings to "hnappur"; there are other denotations assigning one or more of these meanings to nine other lexemes, but none of those lexemes is in Arabic. So, without automated inference, the system cannot translate "hnappur" into Arabic. Simple two-hop translation, namely translation with only one intermediate lexeme, is one kind of inference, though it is susceptible to errors. We reach five Arabic lexemes by translating in two hops from "hnappur". The green disks in Figure 9 represent meanings, and the letters labeling them represent lexical resources. Thus, in this example there are five resources participating in two-hop translations from "hnappur" into Arabic. We are translating through some ambiguous lexemes such as "stud", "key", and "touche", and nothing guarantees that the meanings they share with "hnappur" are equivalent to the meanings they share with Arabic lexemes. But some of the Arabic lexemes have more connections to "hnappur" than others do, and inference routines invented at the Turing Center use such path redundancy as evidence of validity. Three-hop connections provide even more evidence. For example, Esperanto "klavo" = Hungarian "billentyű" = Arabic "مفتاح".
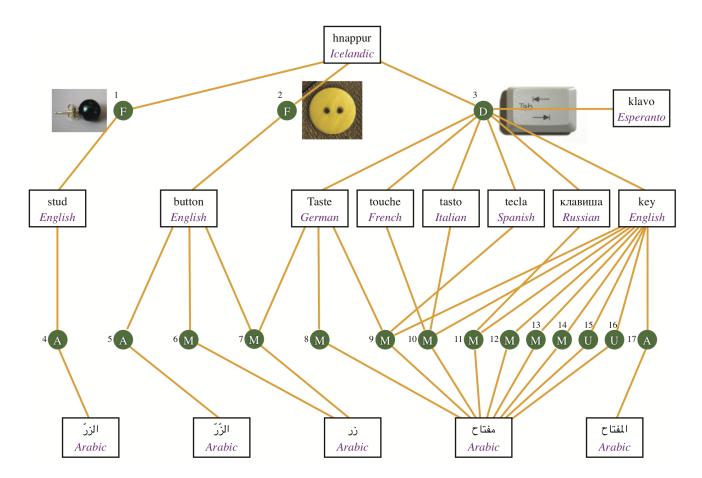
Figure 7. Illustration of the Need for Translation Inference.

Experiments were conducted with inference algorithms applied to an early version of the database, containing about 1,300,000 lexemes (Etzioni 2007). One of the simpler algorithms assumed that any hop on any path exhibits a uniform probability of semantic shift. Another assumption was that cliques (sets of three or more lexemes that are all pairwise translations of each other) have a high probability (about 80 percent, based on tests) of sharing a real meaning. An example of such a clique in Figure 9 is "hnappur", "button", and "Taste", where resource F asserts a shared meaning between "hnappur" and "button", resource D between "hnappur" and "Taste", and resource M between "button" and "Taste". Two algorithms derived inferred translations beyond the attested ones for three language pairs: English-Russian, English-Hebrew, and Turkish-Russian. Persons who were bilingual in these pairs judged the correctness of all the translations without knowing which ones were attested and which were inferred. On average, the judges considered about 92 percent of the *attested* translations correct and about 80 percent of the combined attested and inferred translations correct. With this reduction in precision, the system was able to increase the number of translations by 33 percent for English–Russian, 80 percent for English–Hebrew, and 215 percent for Turkish–Russian.

Inference can also draw on external data. In one set of experiments (Sammer 2007), the attested translations were supplemented with monolingual corpora of news articles. Given an ambiguous lexeme (such as "plant" in English) and translations from it into two other languages, the system determined what fraction of the words found near the target words in the two languages' corpora represented translations of each other. This fraction was positively associated with the lexemes in question sharing a meaning.

Work continues on improved inference algorithms. Initial results on an early version of the data indicate that inference based on redundant paths can expand the sets of translations in a multilingual dictionary by about 50 percent without any increase in error. Given that users reported about 8 percent of attested translations to be erroneous, algorithms that combine translations from multiple sources may be able to discover new translations (increasing "recall") while also eliminating some errors (increasing "precision").

One of the main goals for translation inference is making it efficient. As Figure 7 suggests, a person might easily want an inference algorithm to consider more-than-two-hop paths when extracting translations.

However, experiments conducted at the Turing Center have found multi-hop inference too complex for real-time implementation. Solutions being investigated include precomputation of translations, implementation of the system on clusters of several computers operating in parallel, random sampling instead of exhaustive search for some inference operations, and redefining the problem of translation as a problem of discovering universal meanings and their panlingual expressions. The idea behind this last approach is to discover from the data the real meanings that appear to be most universally expressed in the world's languages and to identify for each meaning an expression in each language. Then users who specify (for instance, with an unambiguous lexeme) one of the universal meanings could obtain its expression in any language instantly, because a time-consuming inference process would not be required.

The **third challenge** has been to show that translations derived from PanLex can produce benefits. The project has pursued this goal by means of two main tactics. One is to show that the translations can make searching the Web more effective, and the other is to show that people can exchange intelligible messages with each other using only translated lexemes.

The search project involved constructing a special Web search engine for images. Launched in September 2007 (Hickey 2007) and made available for public use (http://www.panimages.org), PanImages helps the user formulate and submit multilingual search queries for images. PanImages guides users to type lemmata, helps them choose meanings for the chosen lexemes, and gives them choices among the attested and inferred translations of those lexemes. Users can thereby discover images whose labels are in languages the users don't know, but which are nonetheless relevant to them. The service can also help users (1) to improve the precision of their image-search results by avoiding highly ambiguous query words; and (2) to find culturally specific images (Colowick 2008; Etzioni 2007). PanImages is still an experimental prototype, but it has had about 200,000 visitors in its first year of existence.

A second project investigated lemmatic communication. This is communication in which one person (the "encoder") constructs sequences of lexemes that represent the meaning of a message. An automated system translates the lexemes into another language, and another person (the "decoder") attempts to understand the intended message. The success of this method of communication depends largely on the encoder's avoidance of ambiguous lexemes. For example, in table 3 the encoding and the translation from English to German introduce no major distortion in meaning, so the decoded sentence easily conveys the intended meaning. In table 4, however, ambiguous lexemes in the encoding lead to a translation that describes the last sentence of an essay, instead of the outcome of a criminal case.

Table 3. Example of Successful Lemmatic Communication.

Source Sentence	Encoding	Translation	Decoding
Washing hands regularly is effective in the reduction of the spread of infectious diseases	regularly, wash, hand, effectively, reduce, infectious disease, spread	regelmäßig, waschen, Hand, wirkungsvoll, reduzieren, Infektionskrankheit, ausbreiten	Regelmäßiges Händewaschen reduziert wirkungsvoll die Ausbreitung von Infektionskrankheiten

Table 4. Example of Failed Lemmatic Communication.

Source Sentence	Encoding	Translation	Decoding
The trial ended with a lengthy sentence	trial, end, with, lengthy, sentence	essai, fin, avec, long, phrase	L'essai s'est terminé par une longue phrase

In an experiment on lemmatic communication (Everitt et al. 2010), Spanish- and Hungarian-speaking subjects read passages and converted their sentences to sequences of lexemes. Other subjects read the lexeme sequences and converted them back into passages consisting of sentences. There were three conditions:

1. The lexemes were automatically translated from Spanish into Hungarian or vice versa between the encoding and decoding stages; the translation was crude, always translating a given lexeme identically, regardless of its context.
2. The lexemes were not translated; encoding and decoding subjects spoke the same language.
3. As with condition 2, the lexemes were not translated, but they were randomly reordered—a simulation of word-order differences among languages. The quality of the decoding was rated by another set of subjects.

As expected, both reordering and translation interfered with the task. Still, in all conditions, subjects succeeded in producing final sentences that bore close or moderate resemblance to the original sentences almost half the time or more. On the basis of the subjects' errors and comments, the investigators hypothesized that improvements to the system and user interface could further increase the success of lemmatic communication. The contemplated improvements include more intelligent automatic translation, warnings when encoders choose ambiguous lexemes, options for decoders to see alternate translations, and opportunities for decoders to ask encoders to clarify or try again. The experiment and its pretests revealed that a major issue facing encoders is efficiency. It is difficult to design an encoding interface that allows people to select lexemes from a database as rapidly as they can type free text. However, intelligent interfaces might learn to anticipate the next lexeme and accelerate the selection process, perhaps even exceeding the pace of free-text writing.

Work continues in an effort to make lemmatic communication practical. The Turing Center is developing an application, PanMail, which will allow people to send messages to each other through the internet across all language boundaries, using lemmatic communication. Additional research is under way for designing graphical and other language-independent expressive methods, which can supplement lemmatic communication.

Applications that deliver useful results also create opportunities to collect system-improving knowledge from users. Persons who use systems based on PanLex in order to get translations will sometimes know (or believe) that the translations they get are incorrect, or will be able to perform translations that the system cannot. Experimenting with user-contribution features in the PanImages application, the Turing Center has obtained a few thousand corrections and additions from users. However, these include many jocular, sarcastic, semi-literate, and other low-quality contributions. Obtaining data from many dispersed users requires quality management.

As the PanLex project addresses these three major challenges, its system development can be understood as taking place on three corresponding layers. Layer 1 is the database of attested denotations and auxiliary facts. Layer 2 consists of versions of the database that employ various inference routines developed at the Turing Center for the discovery of unattested translations, universal meanings, and expressions of universal meanings. Layer 3 consists of the applications and experiments that build on the other layers to provide practical services, conduct research, collect additional data, and improve the quality of the existing data.

Future and Related Work

PanLex began as an in-house database for prototypes and experiments designed by one team. Efforts are now under way to move the database and related tools into an institutional and technical environment suitable for easy access to researchers and end-users worldwide. In the envisioned future, the problem of lexical translation inference and the goal of building applications that rely on it will be treated as objects of collaborative and competitive research at multiple institutions. Users anywhere will be able to access the database, add resources to it, and use, evaluate, and improve inference algorithms operating on it. Someone who has constructed a dictionary that translates the words of low-density language A into higher-density language B will, by contributing the dictionary's data to PanLex, enable the speakers of A to translate words from their language not only into B but into thousands of other languages. If this capability, in combination with projects implementing other strategies of panlingual globalization, motivates actions that breathe new life into dying languages, the intuitions underlying PanLex will be shown to have been well-founded.

There appear to be opportunities for mutually beneficial collaboration between PanLex and other projects with similar aims. Collections of digital lexical resources include: Wiktionary (http://www.wiktionary.org/); wordgumbo (http://www.wordgumbo.com/index.htm); FreeLang (http://www.freelang.net/); FreeDict (http://sourceforge.net/project/showfiles.php?group_id=1419); Dicts.info (http://www.dicts.info/); Digital Dictionaries of South Asia (http://dsal.uchicago.edu/dictionaries/); Majstro Aplikaĵoj (http://www.majstro.com/Web/Majstro/sdict.php); Ergane (http://download.travlang.com/Ergane/); Logos (http://www.logos.it/index); OneLook (http://www.onelook.com/); Langtolang (http://www.langtolang.com/); Lingoes (http://www.lingoes.net/en/translator/index.html); jARGOT (http://www.jargot.com/); EUdict (http://www.eudict.com/); SensAgent (http://dictionary.sensagent.com/); OmegaWiki (http://www.omegawiki.org/); WinDictionary (http://www.windictionary.com/); LingvoSoft (http://www.lingvozone.com/); and Webster's Online Dictionary (http://www.websters-online-dictionary.org/). A much larger collection is that of the printed dictionaries in the world's libraries. Projects that digitize books (including dictionaries), such as Project Gutenberg (http://www.gutenberg.org /wiki/Main_Page) and the Google Books Library Project (http://books.google.com/googlebooks /library.html), are other potential content contributors. Relevant standards with which PanLex wholly or

partly complies include Unicode (Unicode 2007) and OLIF (http://www.olif.net/documentation.htm). The Global WordNet Association (http://www.globalwordnet.org/) and Language Grid (http://langrid.nict.go.jp/en/index.html) are other related initiatives.

Mutually beneficial terms of collaboration may be tricky to negotiate with compilers of lexical resources. Many such resources are deployed as advertising-supported services that seek to maximize human visitors in order to generate revenue. PanLex, by contrast, seeks to achieve panlingual *transparency*, in which users get efficient translation without spending time personally choosing and using tools on translation Web sites. The two models might be difficult to reconcile (see Kilgarriff 2000). Moreover, the legal rules under which providers of lexical resources operate are obscure (Zhu et al. 2002; Kienle et al. 2008) and globally unharmonized (Fernández-Molina 2004). There is little relevant case law, and apparently none on lexical resources. Creators of translingual dictionaries sometimes assert claims that their contents are protected by copyright, even while they borrow liberally from other dictionaries on the theory that lemmatic translations, part-of-speech identifications, and other borrowed facts are inherently ineligible for copyright protection. The designers of PanLex hope to avoid disputes while developing forms of mutually rewarding collaboration that facilitate panlingual communication.

Conclusion

Massive linguistic extinction may not be a necessary consequence of globalization. Several strategies are available for making panlingual rather than unilingual globalization a reality. The PanLex project is an attempt to implement one of those strategies. When several such projects have produced results, work can begin to combine them and to study their interactions. Until then, pronouncements on the inevitable demise of the world's languages will be premature.

ACKNOWLEDGEMENTS

REFERENCES

NOTE. Accessibility of all internet resources mentioned below has been confirmed on January 26, 2010.

Abley 2003. Mark Abley, *Spoken Here: Travels Among Threatened Languages*. Boston: Houghton Mifflin, 2003.

Adams 1979. Douglas Adams, *The Hitchhiker's Guide to the Galaxy*. London: Pan Books, 1979.

Ammon 2006. Ulrich Ammon, "Language Conflicts in the European Union". *International Journal of Applied Linguistics*, 16, 2006, 319–338.

Bastardas 2002. Albert Bastardas i Boada, "World Language Policy in the Era of Globalization: Diversity and Intercommunication from the Perspective of 'Complexity'". *Noves SL. Revista de Sociolingüística*, Estiu 2002, 1–9. http://www6.gencat.cat/llengcat/noves/hm02estiu/metodologia/bastardas1_9.htm

Bender 2005. Emily M. Bender and Dan Flickinger, "Rapid Prototyping of Scalable Grammars: Towards Modularity in Extensions to a Language-Independent Core", in *Proceedings of the 2nd International Joint Conference on Natural Language Processing IJCNLP-05 (Posters/Demos)*, 2005, pp. 203–208.

Berners-Lee 2001. Tim Berners-Lee, James Hendler, and Ora Lassila, "The Semantic Web". *Scientific American*, 284:5, 2001, 34–43.

Blanke 1989. Detlev Blanke, "Planned Languages: A Survey of Some of the Main Problems", in *Interlinguistics: Aspects of the Science of Planned Languages*, ed. Klaus Schubert (Berlin: Mouton de Gruyter, 1989), pp. 63–87.

Colowick 2008. Susan M. Colowick, "Multilingual Search with PanImages". *Multilingual*, 19:2, 2008, 61–63. http://turing.cs.washington.edu/PanImMultilingual.pdf

Crystal 2000. David Crystal, *Language Death*. Cambridge, U.K.: Cambridge University Press, 2000.

De Swaan 2004. Abram De Swaan, "Endangered Languages, Sociolinguistics, and Linguistic Sentimentalism". *European Review*, 12, 2004, 567-580.

Dorr 2006. Bonnie J. Dorr, Eduard H. Hovy, and Lori S. Levin, "Machine Translation: Interlingual Methods", in *Encyclopedia of Language & Linguistics*, Second Edition, ed. Keith Brown, volume 7 (Oxford: Elsevier, 2006), pp. 383–394. ftp://ftp.umiacs.umd.edu/pub/bonnie/Interlingual-MT-Dorr-Hovy-Levin.pdf

Eidheim 1969. Harald Eidheim, "When Ethnic Identity Is a Social Stigma", in *Ethnic Groups and Boundaries*, ed. Fredrik Barth (Boston: Little, Brown, 1969), pp. 39–57.

Etzioni 2007. Oren Etzioni, Kobi Reiter, Stephen Soderland, and Marcus Sammer, "Lexical Translation with Application to Image Search on the Web", in *Proceedings of Machine Translation Summit XI*, 2007. http://turing.cs.washington.edu/papers/EtzioniMTSummit07.pdf

Everitt 2010. Katherine Everitt, Christopher Lim, Oren Etzioni, Jonathan Pool, Susan Colowick, and Stephen Soderland, "Evaluating Lemmatic Communication". *trans-kom*, 3, 2010, 70–84. http://www.trans-kom.eu/bd03nr01/trans-kom_03_01_03_Everitt_et_al_Lemmatic_Communication.20100531.pdf.

FAO 2008. Food and Agriculture Organization of the United Nations, *AGROVOC Thesaurus*, http://www.fao.org/agrovoc, 2008.

Ferraro 2002. Paul J. Ferraro and Agnes Kiss, "Direct Payments to Conserve Biodiversity". *Science*, 298, 2002, 1718–1719.

Fernández-Molina 2004. J. Carlos Fernández-Molina, "The Legal Protection of Databases: Current Situation of the International Harmonization Process". *Aslib Proceedings: New Information Perspectives*, 56, 2004, 325–334.

Fettes 2003. Mark Fettes, "The Geostrategies of Interlingualism", ch. 3 in *Languages in a Globalising World*, ed. Jacques Maurais and Michael A. Morris (Cambridge, U.K.: Cambridge University Press, 2003), pp. 37–46.

Gordon 2005. Raymond G. Gordon, Jr. (ed.), *Ethnologue: Languages of the World*, 15th edn. Dallas: SIL International, 2005. http://www.ethnologue.com/.

Harrison 2007. K. David Harrison, *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*. New York: Oxford University Press, 2007.

Hickey 2007. Hannah Hickey, "A rose is a rozsa is a 薔薇: Image-search tool speaks hundreds of languages", http://uwnews.washington.edu/ni/article.asp?articleID=36524, 2007.

Hutchins 2006. John Hutchins, "Machine Translation: History", in *Encyclopedia of Language & Linguistics*, Second Edition, ed. Keith Brown, volume 7 (Oxford: Elsevier, 2006), pp. 375–383. http://www.hutchinsweb.me.uk/EncLangLing-2006.pdf

Jenkins 2004. Michael Jenkins, Sara J. Scherr, and Mira Inbar, "Markets for Biodiversity Services: Potential Roles and Challenges". *Environment*, 46, 2004, 32–42.

Kienle 2008. Holger Kienle, Daniel German, Scott Tilley, and Hausi Müller, "Managing Legal Risks Associated with Intellectual Property on the Web". *International Journal of Business Information Systems*, 3, 2008, 86–106.

Kilgarriff 2000. Adam Kilgarriff, "Business Models for Dictionaries and NLP". *International Journal of Lexicography*, 13, 2000, 107–118.

Libert 2000. Alan Libert, *A Priori Artificial Languages* (München: Lincom Europa, 2000).

Libert 2003. Alan Libert, *Mixed Artificial Languages* (München: Lincom Europa, 2003).

Mufwene 2002. Salikoko S. Mufwene, "Colonization, Globalization and the Plight of 'Weak' Languages". *Journal of Linguistics*, 38, 2002, 375–395.

Mufwene 2004. Salikoko S. Mufwene, "Language Birth and Death". *Annual Review of Anthropology*, 33,

2004, 201–222.

Nettle 2000. Daniel Nettle and Suzanne Romaine, *Vanishing Voices: The Extinction of the World's Languages* (Oxford: Oxford University Press, 2000).

Panfilov 2008. Кирилл Панфилов, "Баскско-русский словарь", http://www.erlang.com.ru/euskara/?basque-eurus, 2008.

Phillipson 2008. Robert Phillipson, "*Lingua Franca* or *Lingua Frankensteinia*? English in European Integration and Globalisation". *World Englishes*, 27, 2008, 250–284.

Pool 1991. Jonathan Pool, "The Official Language Problem". *American Political Science Review*, 85, 1991, 495–514.

Pool 2006. Jonathan Pool, "Can Controlled Languages Scale to the Web?", in *Proceedings of the 5th International Workshop on Controlled Language Applications (CLAW 2006)*, 2006. http://turing.cs.washington.edu/papers/pool-clweb.pdf.

Raverty 1867. Raverty, Henry George, *A Dictionary of the Puk'hto, Pus'hto, or Language of the Afghans*, 2nd edn. (London: Williams and Norgate, 1867). http://dsal.uchicago.edu/dictionaries/raverty/index.html

Sammer 2007. Marcus Sammer and Stephen Soderland, "Building a Sense-Distinguished Multilingual Lexicon from Monolingual Corpora and Bilingual Lexicons", in *Proceedings of Machine Translation Summit XI*, 2007. http://turing.cs.washington.edu/papers/SammerMTSummit07.pdf.

Schubert 1992. Klaus Schubert, "Esperanto as an Intermediate Language for Machine Translation", in *Computers in Translation*, ed. John Newton (London: Routledge, 2006), pp. 78–95.

SIL 2008. SIL International, *ISO 639-3*, http://www.sil.org/iso639-3/default.asp, 2008.

Tonkin 2003. Humphrey Tonkin, "The Search for a Global Linguistic Strategy", ch. 21 in *Languages in a Globalising World*, ed. Jacques Maurais and Michael A. Morris (Cambridge, U.K.: Cambridge University Press, 2003), pp. 319-333.

Trujillo 1999. Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*. London: Springer, 1999.

UDHR 2008. Unicode Consortium, "UDHR in Unicode", http://www.unicode.org/udhr/, 2008.

UNESCO 2003. UNESCO Ad Hoc Expert Group on Endangered Languages, "Language Vitality and Endangerment", International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages, 2003. http://www.unesco.org/culture/ich/doc/src/00120-EN.pdf.

Unicode 2007. Unicode Consortium, *The Unicode 5.0 Standard*. Upper Saddle River, N.J., U.S.A.: Addison-Wesley, 2007.

Van Parijs 2007. Philippe Van Parijs, "Tackling the Anglophones' free ride: Fair linguistic cooperation with a global lingua franca". *AILA Review*, 20, 2007, 72–86.

Wikimedia Foundation 2008. Wikimedia Foundation, *Wiktionary*, http://www.wiktionary.org, 2008.

Woodbury 2006. Anthony C. Woodbury, "What is an Endangered Language?". Linguistic Society of America, 2006. http://www.lsadc.org/info/pdf_files/Endangered_Languages.pdf.

Zhu 2002. Hongwei Zhu, Stuart E. Madnick, and Michael D. Siegel, "The Interplay of Web Aggregation and Regulations". *Proceedings of Law and Technology, LAWTECH 2002*, 2002, Track 375-853.