Breaking Language Barriers: Dream or Reality?

Professor Dr. Alexander Waibel, Karlsruhe

Abstract

In an increasingly interconnected world, the inhabitants of our planet are no longer separated by the lack of digital communication technology (digital divide), but by language barriers. 80% of the world's population already have a mobile phone, via which everyone can reach nearly everyone. But with more than 6000 languages, not to mention numerous dialects and accents, we can understand only few of our fellow humans. Human interpreters alone – even though they are qualitatively better – cannot cover this growing demand for communication. Technical aids are indispensable. Automatic translation not only of text, but also of spoken language is now becoming possible: Since the late 1980s, our research has been devoted to this challenge. Initial prototypes were still slow, domain-limited, and inflexible. In the course of time, they have developed to practically deployable systems for everyday communication today. Mobile pocket interpreters for tourists, dialog interpreters for doctors and humanitarian organizations, automatic real-time simultaneous translators for lectures, TV news, or speeches are made available as smartphone apps or cloud-based services.

The present article outlines scientific challenges associated with computer interpretation of spoken language, describes the technologies used for this purpose, and discusses the development phases and uses of today's cross-lingual language communication systems and services.

Introduction

In an increasingly interconnected world, digital communication has long ceased from being the biggest barrier separating human beings. The digitized world is interconnected throughout and even in remotest parts of our world, nearly every inhabitant can be reached by mobile phone. More than 6 billion mobile phones are presently in use, one phone per global citizen on the average. Developing countries in particular consider the connection to the telephone network and internet a prerequisite for a better future and economic survival. Of course, many regions of the world are not yet accessible by internet and telephones. But decreasing costs and the wish of individuals to participate in the network push this process with high speed. Even at remotest places in the world can communication antennas be set up quickly and is it possible to speak with any other person on this planet via satellite radio or telephone. The internet giants Google, Facebook, and others are presently competing for being the first to open up these remote regions of our world with drones, balloons, and similar means.

In view of this situation and with rapidly increasing globalization, the so-called “digital divide” is being replaced by the “linguistic divide” as the bigger one of the communication problems between human beings. The existing Babylonian confusion of languages continues to preserve communication barriers between us: We may reach every person on our planet, but we do not understand him or her. About 6000 to 7000 languages exist on earth. Consequently, more than 36 million translation directions would be needed for everyone to communicate with everyone. For every language pair in the European Union (with 24 official languages) already, there are not always translators, who know both languages. This problem is mainly addressed by three solutions:

1. Learning more languages.
2. Using English (or Latin?, Chinese?, …) as common lingua franca.
3. Using human translators or interpreters for communicating.

The first solution is important and culturally healthy, but rather bothersome. As a rule, only few language pairs can be mastered personally and individually. Multilingual speakers, who fluently speak three, four, or even five languages, are fortunate, but rare.

The second solution of everybody learning English is unrealistic and culturally questionable: Should human beings and nations give up their cultural diversity, individuality, and the singularity of their language? Should a country be forced in the long term to read its own literature in translations only? Aside from cultural loss, this solution also is unrealistic. On which language should we agree? Why should we agree on English? Why not on French, Spanish, Chinese or Latin? And how can we ensure the same linguistic competence for everybody and prevent social barriers from being caused by linguistic ones. In Europe alone, where language education is promoted comparably well and mandatory at schools, an average of only 34% of all Europeans know enough English to be able to effectively work with it.

The third solution of using human interpreters and translators unfortunately has to be rejected, as it is impractical: With the exception of a few critical areas where this is possible and important (the European Parliament, for instance), it would be unaffordable for large parts of society.

It is therefore necessary to find a fourth solution and to effectively integrate it into our communication paths: An automatic, computer-based, inexpensive or cost-free solution. Translingual communication, however, must not be understood to be the translation of texts only. Language is spoken, written, composed, and drawn. Hence, language first has to be understood in its original form of expression before the attempt can be made to translate it.

Technology

This topic is addressed by a growing community of scientists. Different components handling partial aspects of the communication problem have to be distinguished (see figure). For a

human being speaking one language to understand another human being speaking another language, three partial tasks have to be solved.

1.) Automatic speech recognition: Here, the signal spoken in language 1 (L_a) is recorded by microphone, processed, and output as text (speech to text), 2.) machine translation: Here, text in one language (L_a) is translated into text in the other language (L_b) (text to text), and 3.) speech synthesis (L_b): Here, text in the target language L_b is output in spoken language (text to speech). For a dialog between persons speaking two languages, this process also has to be possible in the other direction (from b to a) and, hence, requires analog subsystems in the other language. A final integration of these subsystems with a comfortable user interface then has to be operable easily in real communication situations.

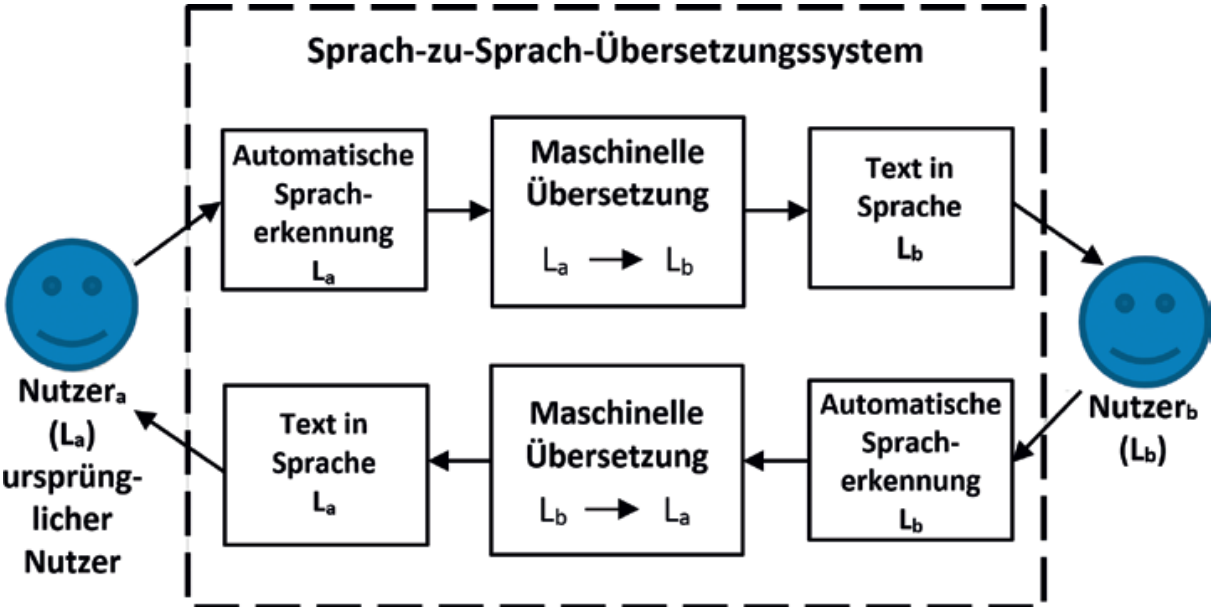

Figure 1: Translation of spoken language (speech to speech translation) – overview.

Each of these partial tasks represents an area of research, which has turned out to be much more difficult than initially expected due to the complexity and ambiguity of human language. For this reason, they have been studied by scientists for several decades already and are still challenging them in spite of the big progress achieved. The most important lessons learned are that a.) we can never make hard, but only soft probabilistic statements for every source of knowledge in human language due to its ambiguity and b.) we cannot encode these statements and their interactions manually due to their complexity, but can only learn them from data.

Automatic Speech Recognition (ASR)

For the unaware observer, the problem of speech recognition may not appear very difficult at first, as we human beings manage it well and easily. However, several ambiguities occur in spoken language already: The English acoustic sequence of sounds/ juθəneʒə/ (in phonetic transcription) may mean both “Euthenasia“ or “youth in Asia“. Sentences like “This machine can recognize speech“ are pronounced in the same way as “This machine can wreck a nice beach“. Speech recognition requires an interpretation as to which of several similar alternatives is the more meaningful or more probable one in a given context. In modern speech recognition systems, this is achieved by a combination of acoustic models that assign a probability to every sound, a pronouncing dictionary (that assigns a pronunciation to every word), and a language model that evaluates the probability of every possible word sequence “w1, w2, …” of the sentence. Figure 2 shows such a typical decoder. Evaluation of these models during recognition and settings of the best parameters of these models, however, cannot be determined manually, but require automatic search and optimization algorithms.

Parameters of acoustic and linguistic models are learned with the help of learning algorithms using huge databases of speech samples, whose transcriptions are known.

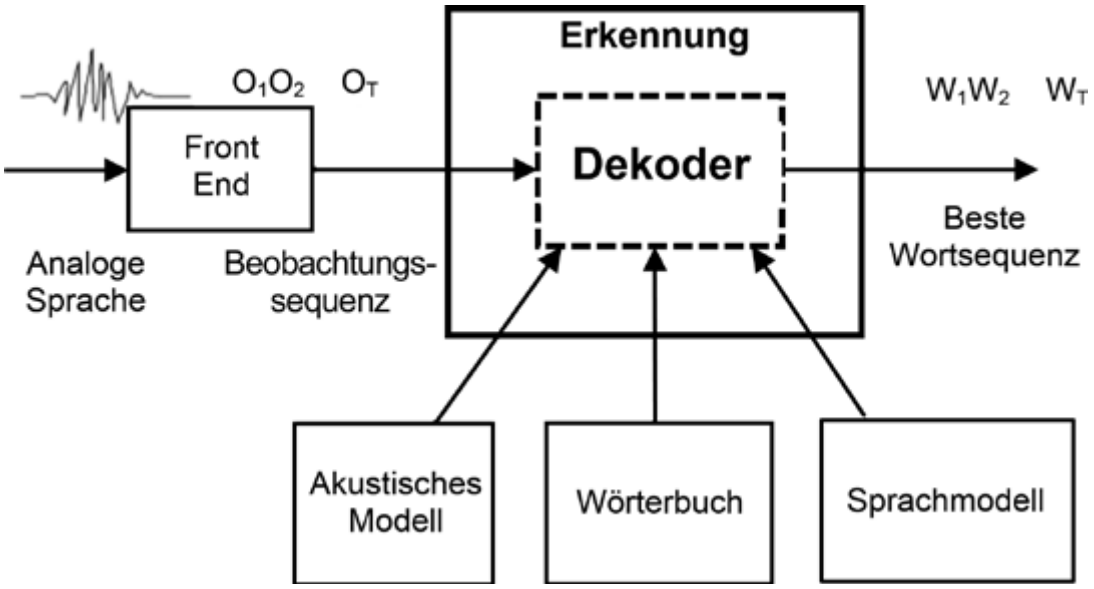

Figure 2: A typical speech decoder (speech to text).

Algorithms work with statistical optimization methods or neural networks and learn the best match between signals and symbols (context-depending phonemes and words) based on known exemplary data. Today's systems use neural networks with several millions of neural links optimized by the learning algorithm.

Machine Translation (MT)

First attempts to translate texts by machines (MT = machine translation) were made as early as during the world wars, but these and later attempts failed due to the ambiguity of language and the complexity of resolving it with the help of associated context knowledge. Nearly every word (skate, row, mouth) has several meanings and, hence, translations that can only be interpreted correctly in the context. The sentence from the bible "The spirit is willing but the flesh is weak" is reported to have been translated into Russian by a first machine translator as "The vodka is good but the flesh is rotten" (old times of MT). And also language structure may be ambiguous. For instance, what does the pronoun "it" refer to in "If the baby doesn't like the milk, boil it"? Of course, the author means that the milk has to be boiled, not the baby!

The attempt to manually encode the required syntactic, semantic, and lexical knowledge with the help of rules failed again and eventually had to give way to automatic learning processes after decades of research. Now, a modern MT system uses a similar system architecture that optimally combines a several learned statistic knowledge components (see Fig. 3).

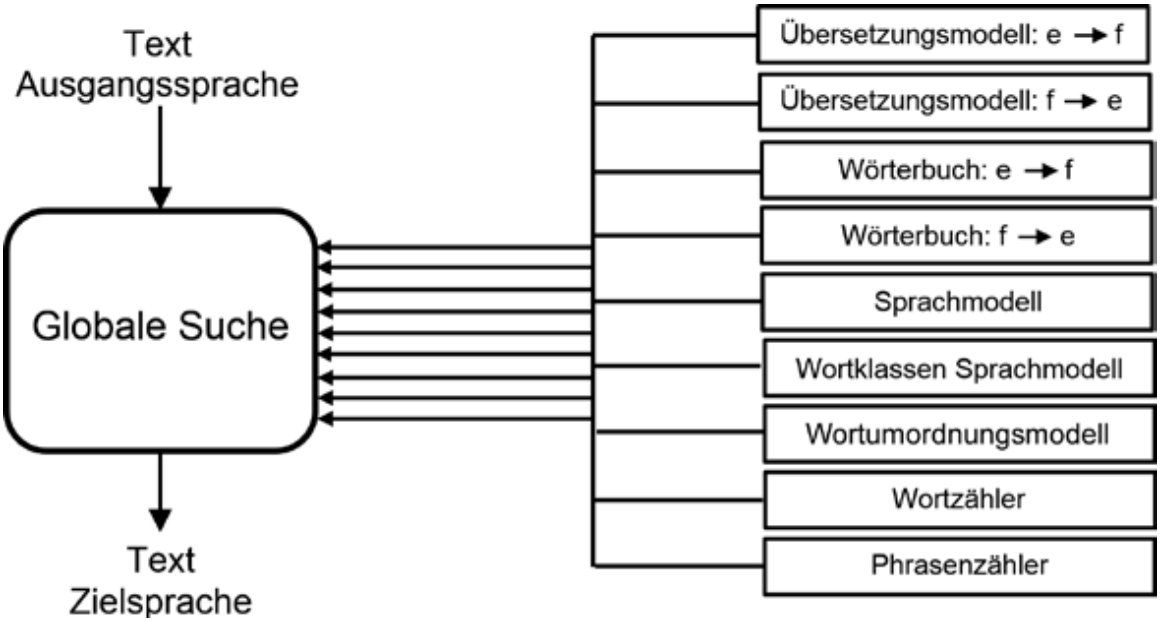

Figure 3: Machine translation (text to text).

Speech Synthesis (Text to Speech (TTS))

The third component usually is speech synthesis. It is to make the translated sentence audible in the target language, i.e. to speak it. In comparison, TTS synthesis may be the simpler component, as exactly one signal is to be produced by a given textual sentence and it is not necessary to handle the comparably many ambiguities of the other components. But again, pitfalls exist that make the problem continue to be a subject of research. For example, how can text be transcribed into phonetics (e.g. though → "th oh"). And how are these allocations to be made in many languages, in which pronunciation rules differ considerably? Again, automatic learning algorithms are used to optimize settings and matches based on prefabricated or existing training material.

Evolution of Systems

	Jahre	Vokabular	Sprechstil	Domäne	Geschwindigkeit	Plattform	Example Systems
First Dialog Demonstration Systems	1989-1993	Restricted	Constrained	Limited	2-10 x RT	Workstation	JANUS-1, C-STAR-I
One-Way Phrasebooks	1997-Present	Restricted, Modifiable	Constrained	Limited	1-3 x RT	Handheld	Phraselator, Ectaco
Spontaneous Two-way Systems	1993-Present	Unrestricted	Spontaneous	Limited	1-5 x RT	PC/ Handheld Devices	JANUS-III, C-STAR, Verbmobil, Nespole, Babylon, Transtac
Translation of Broadcast News, PoliticalSpeeches	2003-Present	Unrestricted	Read/ Prepared Speech	Open	Offline	PC's, PC-Clusters	NSF-STRDUST, EC TC-STAR, DARPA GALE,
Simultaneous Translation of Lectures	2005-Present	Unrestricted	Spontaneous	Open	Realtime	PC, Laptop	Lecture Translator
Commercial Consecutive Translators on Phone	2009-Present	Unrestricted	Spontaneous	Open	Online and Offline	Smartphone	Jibbigo, Google, Microsoft,
Simultaneous Interpretation *Services*	2012-Present	Unrestricted	Spontaneous	Open	Online	Server, Cloud-Based	KIT, EU-Bridge, Microsoft

Table 1: Development phases of speech translation systems.

Development of automatic translation systems of spoken language started in the early 1990s when first ASR, MT, and TTS systems reached the minimum degree of maturity required for first integration. In the course of the following two decades, major limitations of technology were overcome in a number of research and development phases. Today, speech translators can be used (see Table 1 for an overview of system qualifications).

First Demonstrators

The JANUS system was the first speech translation system presented to the public in the USA and Europe in 1991. JANUS was developed for German, Japanese, and English by Universität Karlsruhe and Carnegie Mellon University of Pittsburgh, USA. It was a result of cooperation with the ATR Interpreting Telephony Laboratories in Japan, which developed similar systems for the Japanese language in parallel. The systems together were presented in first translating video conferences (Waibel et al., 1991, Handelsblatt, July 30, 1991).

These systems represented first steps, managed an initially small vocabulary (< 1000 words), required a relatively restricted syntax, and covered a limited domain (e.g. registration for a conference). They were too large and slow to really be of assistance in field situations, e.g. to a traveler. Similar demonstration systems were presented by other research groups (ATT and NEC) in 1992.

Figure 4: First speech translation prototypes in video conferences (1991).

Research Systems and Prototypes

For these systems to be used in practice, other important phases of development followed to successively master difficult problems:

Spontaneous speech, domain-limited research systems: To implement practical systems, the assumption of syntactic correctness has to be eased or eliminated. People rarely speak syntactically correct and complete sentences. They rather speak fragmentary scraps with stammerings, repetitions, filler words, and breaks (errs, emms, etc.). These fragments first have to be identified correctly and then filtered out or corrected by processing before translation takes place. First spontaneous speech translation systems were developed from 1993 to 2000 (Morimoto et al., 1993, Takezawa et al., 1998). These systems were still slow and required extensive hardware. Their domain continued to be too limited to extract the fragments relevant to translation by modeling the semantics. JANUS-III, C-STAR Systems, VERBMOBIL, and other projects achieved progress, but first remained unusable in practice (Lavin et al., 1997). Domain limitation had to be eliminated and the systems had to be made more rapid and mobile for use. Manually programmed rules (possible in limited domains) were replaced by automatically learned, statistic subsystems in order to improve robustness and scalability (Brown et al., 1993, Och and Ney, 2004, Koehn at al., 2007).

Two different applications began to emerge, which were associated with additional technical challenges:

- Consecutive interpretation of dialogs in mobile use: Here, a dialog with a partner speaking another language is to be enabled. Sentences of both conversation partners are translated first for the respective partner to reply. As a rule, every conversation partner speaks one to two sentences for consecutive interpretation. In most applications (tourism, medical uses, humanitarian aid, …) a general vocabulary of about 40,000 words is required only. But translation has to be provided rapidly (to prevent conversation flow from stopping) and the system has to be mobile (on a portable small hardware).

- Simultaneous interpretation for stationary use: In many areas of speech translation, no dialog between two conversation partners, but the quick, effective translation of a monolog is needed. Examples are TV broadcasts, internet videos, lectures, speeches, and addresses. Mobility mostly is of no relevance to these applications, as computations can be made online or offline on big servers. Often, more complex issues with technical terms and technical jargon have to be translated. And the speaker does not only produce one or two sentences, but a speech, i.e. a continuous flow of words. The system itself has to find the start and end of translatable units or sentences. Such a segmentation into units or fragments and automatic punctuation (full stops, commas, question marks) have to be made automatically taking into account the context (The Economist, June 12, 2006).

Translation of Speech in Practice

Early research systems (1990 – 2005) solved technical problems and paved the way for the sales and real use of speech translation systems in society.

Consecutive Interpretation

Interpretation systems were first tested in the field during humanitarian and logistic exercises of the US government. Initially, domain-limited speech-to-speech translation systems were used (MOBILE TECHNOLOGIES) (Eck et al., 2010). As an alternative, simple language-based phrase books (VOXTEC) were applied to retrieve prefabricated phrases via speech input without translation components (http://www.voxtec.com). Early models of these commercial systems by VOXTEC and MOBILE TECHNOLOGIES are shown in Figure 5. However, these systems were domain-limited in vocabulary and language use and, hence, designed for dialogs in special areas only (use in crisis areas, medical missions, police, hotel reception, etc.). Sales and commercialization were limited to small user groups.

With the introduction of smartphones, computation capacity of mobile phones reached the critical capacities for speech recognition and translation of open unlimited (> 40,000 words) vocabularies in nearly real time.

Figure 5: First commercial systems: Phraselator, iPaq PDA-based speech translator[1] (~ 2005), and (right) JIBBIGO, the world's first speech-to-speech translator on a phone (2009).

MOBILE TECHNOLOGIES (a startup of the Carnegie Mellon Laboratories) commercialized the world's first domain-unlimited speech-to-speech translation system on a phone in 2009: JIBBIGO (Eck et al., 2010). Inspired by the simple sales mechanisms of the Apple iTunes app stores and the growing use of smartphones worldwide, JIBBIGO quickly expanded to 15 languages and reached wide worldwide dissemination. JIBBIGO opened the market for portable speech translators. Alternative products were commercialized by Google and Microsoft. Initially, speech processing took place via the internet on external servers. Hence, these translators could be used in case of an existing internet connection only.

Mobility, low costs, large vocabularies, and general availability (also without network connection) of an offline solution, with all components running locally on the phone, are independent of the internet and, hence, more useful for travelers (no roaming costs) and humanitarian missions (no internet required). Consequently, JIBBIGO was the platform

[1] Phraselator by VOXTEC LLC and Speech Translator by Mobile Technologies LLC.

preferably used in a number of humanitarian missions of the US government and charitable NGOs in Thailand, Cambodia, and Honduras for translation between English-speaking physicians and patients speaking other languages (see Figure 6A-D).

Figure 6: Medical operations in Thailand, Cambodia and Honduras: (A)Translingual dialogs between American physicians and patients in Thailand, (B) Medical care with help of the JIBBIGO-speech to speech translator in Thailand, -
(C) Medical operation in Cambodia and (D) Humanitarian operations with JIBBIGO in Honduras.

The systems may be used on iPhone or Android telephones, but tablet computers were found to be particularly suited for user-friendly interaction of partners sitting opposite to each other in humanitarian missions.

After five years of development in field situations, the systems were also evaluated in humanitarian missions (Medical Civil Action Program) in Thailand in 2013 (Gov. Report OMB No.0704-0188, "Speech-to-Speech Translation Tool (S2T2) Limited Utility Assessment Report", Scott Hourin et al., 2013). 95% of the interactions associated with the registration of patients were managed with the help of an automatic JIBBIGO tablet interpreter without a human interpreter being required.

Simultaneous Interpretation

In a multilingual environment, dialog between conversation partners speaking different languages is not the only challenge. When thinking of TV news, films, presentations, lectures, speeches, road signs, transparencies for lectures, and short messages, we see many other challenges, where translingual technologies are required.

An important area of application is the translation of lectures. In spite of excellent scientific equipment and funding, German universities in particular are often disadvantaged in international competition for talents, simply because many foreign students or scientific employees and academics do not want to learn another new language (especially such a difficult language as German). How are German universities or German companies to react? Is a German university supposed to have all courses and lectures presented in English? The author of this article does not consider this desirable or practicable. A hybrid solution with the help of modern language technologies that supports linguistic and cultural diversity and tolerance (and does not suppress one or the other direction) appears to be far more promising, as it fosters and improves internationalization and international understanding.

At Karlsruhe Institute of Technology (KIT), such a system is being used for students in the main auditorium (Cho et al., 2013). Speech translation continues to be the subject of research, as not all problems are solved. But for a listener, who does not speak the language of the lecturer, an imperfect computer-based interpreter is better than nothing.

The system first presented by CMU and KIT in 2005 (see Figure 7) manages one direction initially, as a lecture is a monolog to be translated from one into the other language. Such a system does not have to be run on a mobile and portable device, but may be operated on servers in a cloud-based manner and accessed via the internet. Contrary to a dialog system, a lecture translator requires a speech recognition component and a translation machine. Speech synthesis can take place afterwards, but it is optional, if subtitles are desired. In addition, a segmentation component is required to decide explicitly or implicitly when the

end of a sentence or at least of a translatable fragment is reached in the flood of words. Vocabularies containing many technical terms and jargon may be a problem during lectures.

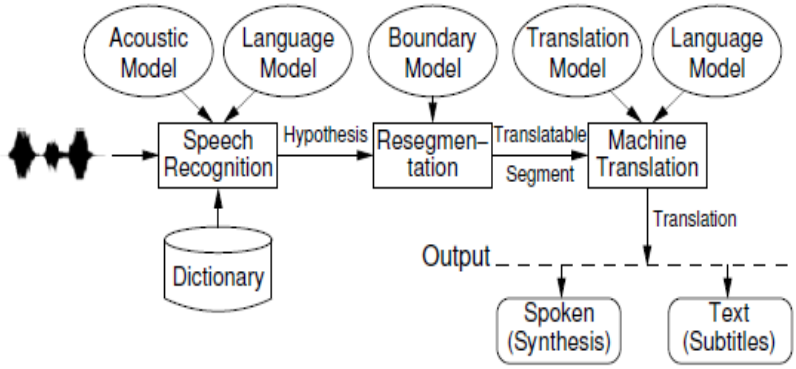

Figure 7: Speech translation of lectures.

A lecture translator may be operated online or offline. An online system is required when a listener wishes to follow subtitles (during the lecture) in one language (transcript) or both languages (translation). Online use requires real-time recognition and translation (i.e. the system has to keep up with the speech). Latency (i.e. the time lag between the spoken word and the translated word) is to be minimized. Otherwise, the listener will lose track of the lecture and of what is happening in the lecture hall. Especially in the German language, these requirements are a challenge, because the verb or important parts of the verb sometimes appear at the end of the sentence (or sometimes even later) only. In a sentence like "**Ich schlage** Ihnen nach eingehender Prüfung Ihres Antrags, der uns gestern …. und … nachdem … und …. eine neue Vorgehensweise vor", the small word "vor" that sometimes appears after several minutes only decides on whether the English sentence starts with "**I propose …"** or **"I hit …".**

In many application scenarios of academic teaching and multimedia broadcasting, offline processing of speech and translation are acceptable or desirable. Offline operation does not necessarily require real-time capability (although an excessively long processing time may become a relevant cost factor) and the system can make a better transcription and translation when taking into account a longer context. A lecture translator, for instance, may

be run online in the lecture hall and then the output may be reprocessed in the offline mode for storing an improved version for the listener in the archive.

Such a lecture translation system was taken into operation by KIT as an internet service in the main auditorium in 2012 (Fig. 8) (Cho et al., 2013, Spiegel Online, June 12, 2012). Students, who wish to have translation support, connect their phones, tablets or PCs to a specially generated website via a normal internet web browser and are provided with a simultaneous transcription of the text in German (useful in case of hearing problems) and a translation into English. Other output languages are under development at the moment.

Figure 8: The lecture translator in use in the main auditorium of KIT.

Transcription and translation of lectures at universities are associated with other, still unsolved problems that are covered by advanced research, in particular in the German language. In addition to the problems of word order and verbs discussed above, the following difficulties are encountered in the German language:

- Compound words: German words like "Fehlerstromschutzschalterprüfung" first have to be decomposed before they can be translated into English. Our Institute develops algorithms to decompose compound words. But due to the ambiguity of language, this is not always easy. Decomposition into *Fehler-Strom-Schutz-Schalter-Prüfung* in our example is reasonable, but decomposition of "dramatisch" into "Drama-Tisch" or of "Asiatisch" into "Asia-Tisch" may be inappropriate in the context or even change the meaning (Koehn and Knight, 2003).
- "Agreement": Suffixes in the German language have to be consistent and fit to the nouns: "in **der** wichtig**en**, interessant**en**, didaktisch gut vorbereitet**en**, heute und gestern wiederholt stattfindend**en Vorlesung**".
- Technical terms and jargon: This is a big problem, in particular when processing lectures at a university, because every lecture has its own technical terms and linguistic features. What are "Cepstral-Koeffizienten" (cepstral coefficients), "Walzrollenlager" (roller bearings), and "Würfelkalküle" (cube calculi), etc.? For a partial solution, we use automatic algorithms that search the transparencies of the lecturer for unknown words and automatically include them (and related technical terms found on the internet) in the recognition vocabulary. As an alternative, we also use the feedback of students and their spontaneous corrections to automatically learn new terms. Then, our programs automatically search for the translations of these technical terms in public internet sources, such as Wikipedia (Niehues and Waibel, 2012).
- "Code switching": Often, lectures and speeches contain quotations and terms in other languages. Especially computer science lectures often contain English terms that are declined in German. Germans talk about the "iPhone", "iPad", "cloud-

basiertem Webcastzugriff" or "Files, die man downgeloaded hat". In German spoken language, English words are sometimes pronounced according to German, sometimes to English rules, they are declined in German, and sometimes even packed into compound words!

- Pronouns: What do pronouns refer to? Here, problems occur rather frequently. The spoken word version of "Wir freuen uns, **Sie** heute hier begrüßen zu dürfen" may be translated as "We are happy to welcome **her** here" or "We are happy to welcome **you** here" (in writing, Germans use a capital and a small "s" to distinguish both versions).
- Readability: When people speak, they do not speak punctuation marks or the ends or starts of paragraphs contained in readable text. Hence, full stops, commas, question marks, paragraphs, and sometimes even titles have to be generated and inserted automatically (Cho et al., 2014).
- Spontaneous speech: Different speakers speak more or less syntactically correctly. Hesitations, stutterings, repetitions, and discontinuations of speech aggravate readability and make translation difficult. A spoken sentence of a lecture in German that is transcribed by a perfect speech recognition system would look like this without any punctuation marks and corrections of spontaneous effects: *„Das ist alles was Sie das haben Sie alles gelernt ja und jetzt können Sie es einsetzen und ich erzähle gleich welche Implikationen das hat Ähm das ist auch so ja und äh wenn Sie die Systeme die Sie bauen dann eben auch einsetzen dann sind Sie wenn Sie so wollen ja der der der erste Tablet ähm den es so gab ähm hatte ein Wireless LAN Ähm eine Charakteristik ist dass wir versuchen unsere ähm den den den Zweck und die Funktionalität zu äh einzuschränken ja da gibt's da gab's in äh gab's nur eines".*

  Again, the speech recognized first has to be processed linguistically in order to make it readable in the source language. Afterwards, it can be translated into readable text in the target language [Cho et al., 2014b].
- Microphones and noise: Our lecture translator presently is configured such that the speaker carries a dynamic noise-canceling microphone. This is acceptable during lecturing, as lecturers carry microphones in auditoriums anyway. In seminars and meetings, however, this would be disturbing. Unfortunately, open table microphones

cause echoes and noise. Speaking of several speakers leads to considerable losses of recognition performance.

- Linguistic scalability/portability: How can we implement the technologies developed not only in one or two languages, but extend it to cover communication among all languages and cultures on our planet? To achieve this, development costs of a translation system would have to be reduced considerably. Language-independent technologies (presently under development with the help of neural networks by our Institute), adaptation, inference, abstraction, better use of monolingual resources, and crowd sourcing (to better harvest the multilingual knowledge of mankind) are promising approaches.

The architecture of the KIT Lecture Translator for practical use was introduced and tested at KIT under the EU-BRIDGE integrated project of the European Union. Now, several lectures can be supported in a cloud-based manner at the same time.

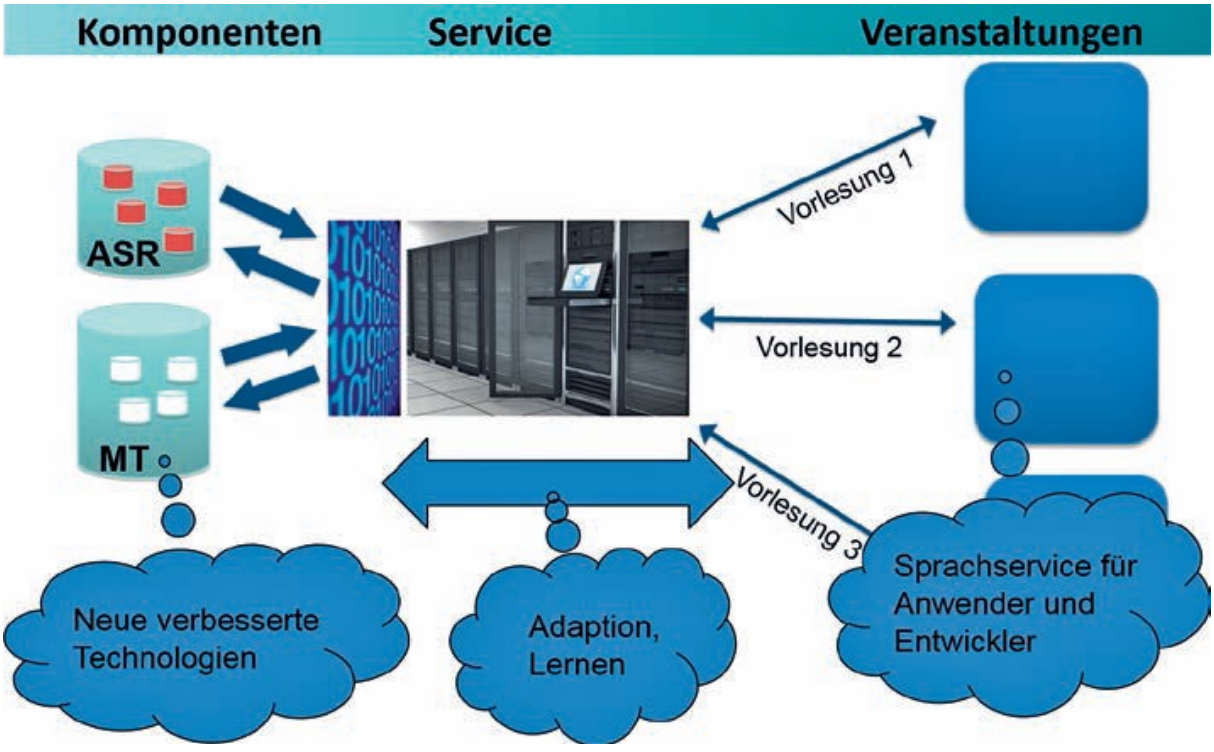

Figure 9: EU-Bridge: The automatic interpreter as a cloud-based service.

Figure 10: Automatically translated speech at the European Parliament.

By means of this server architecture, translation services can be used in several auditoriums and other application scenarios (not only during lectures at university). In 2012, 2013, and 2014, the lecture translation system was presented at the European Parliament, at several Rectors' Conferences, and in training courses of interpreters.

Apart from use at universities (where usually no translation support is provided), automatic systems can also be applied to support experts, for example human interpreters at parliaments. A first test in interpreting booths of the European Parliament was carried out successfully in late 2014 under the EC Integrated Project "EU-Bridge" in Strasbourg.

Figure 11: First test of an automatic interpreter during voting at parliament.

As the European Parliament provides far better interpretation services by human experts (here, the world's best interpreters work with more languages than in any other organization of the world), such services will rather be used as back-up solutions to facilitate the work of interpreters. The technology can automatically generate terminology lists or record figures and names that may then be used as a scratch pad or memory aid by the interpreter. In addition, an "Interpreter's Cruise Control" is conceivable. It might be switched on in case of uncritical or repetitive fragments of meetings (e.g. readout of voting results).

Translingual Communication

In a multilingual and multicultural environment, language barriers are not only encountered in spoken dialogs, lectures, or text documents. They occur in many other communication situations, circumstances, and media: Important information can be found on road signs or in short text messages (SMS), TV news, lecture transparencies, gestures, and many more. To make the vision of a multilingual, language barrier-free world come true, our efforts have to go beyond the construction of better translation systems. We have to try to improve user interfaces that make these language barriers transparent or move into the background. Successful translingual communication is achieved, if people can interact with each other without being aware of barriers.

At the InterACT laboratories of CMU and KIT, we have been working on multimodal user interfaces for transparent communication in various situations parallel to translation for a longer time now. A number of prototypes and use scenarios have already been studied:

- A road sign translator: As early as in 2001 [Yang et al., 2001], systems were presented to read and translate road signs with the help of a mobile camera. Translations were inserted into the image of the scene and the system was tested first on a (then applicable) PDA platform. Meanwhile, similar applications have been developed and issued as apps for iPhones.

Figure 12: Road sign translator.

- Speech translation goggles: In 2005, our simultaneous translations were also output in heads-up display goggles. The user sits opposite a conversation partner and is displayed the translations of spoken words in the form of subtitles in the glasses. In 2005, this still appeared to be science fiction. However, such configurations can be realized and are partly already available on smartphones, smart watches, and Google glasses.

- Handwriting recognizers recognize the handwriting and provide the translation. This problem has been solved partly by the road sign translator or scanner, but still requires easier operation for real environments.
- Translation of lecture transparencies: Our foreign students fail to understand not only the lectures in a foreign language, but also the transparencies of the lecturers. For this reason, a translation system was developed for transparencies at the Institute. It translates the text written on a transparency, if the mouse is moved close to it. The translated text is displayed in a speech bubble.

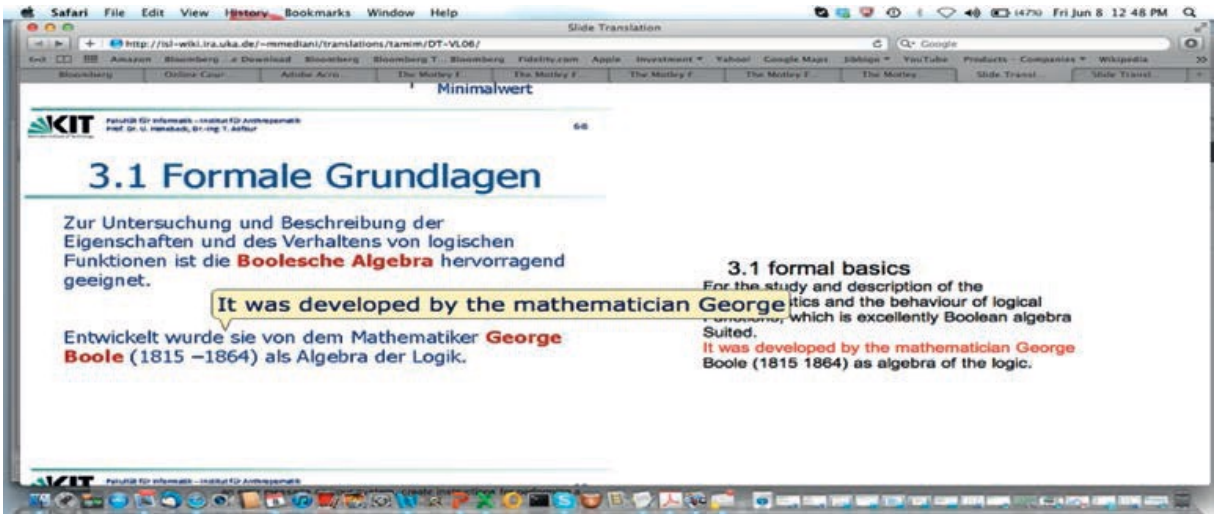

Figure 13: Translation of lecture transparencies.

- Silent speech: Speech always is associated with noise. We also developed a system that recognizes articulated mouth movements by electromyography, even though the language is not spoken out loud. Such silent speech can be recognized (although recognition is not as good as for spoken language), translated, and made audible by synthesis [Maier-Hein et al., 2005]. Consequently, articulation of silent speech can be translated into audible speech in another language.
- Speech translation with the help of directional microphones: With the help of microphones directed accordingly, it is possible to transmit the translation result in the form of synthesized speech to certain points in a room only. In this way, audible simultaneous translation can be presented to individual listeners in various languages even without a headset.

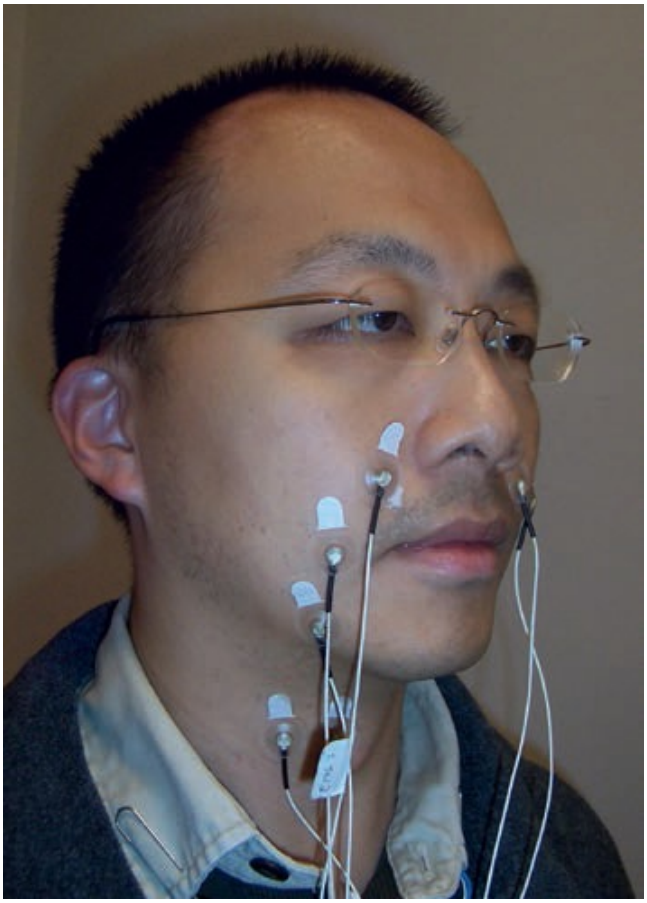

Figure 14: Silent speech translator.

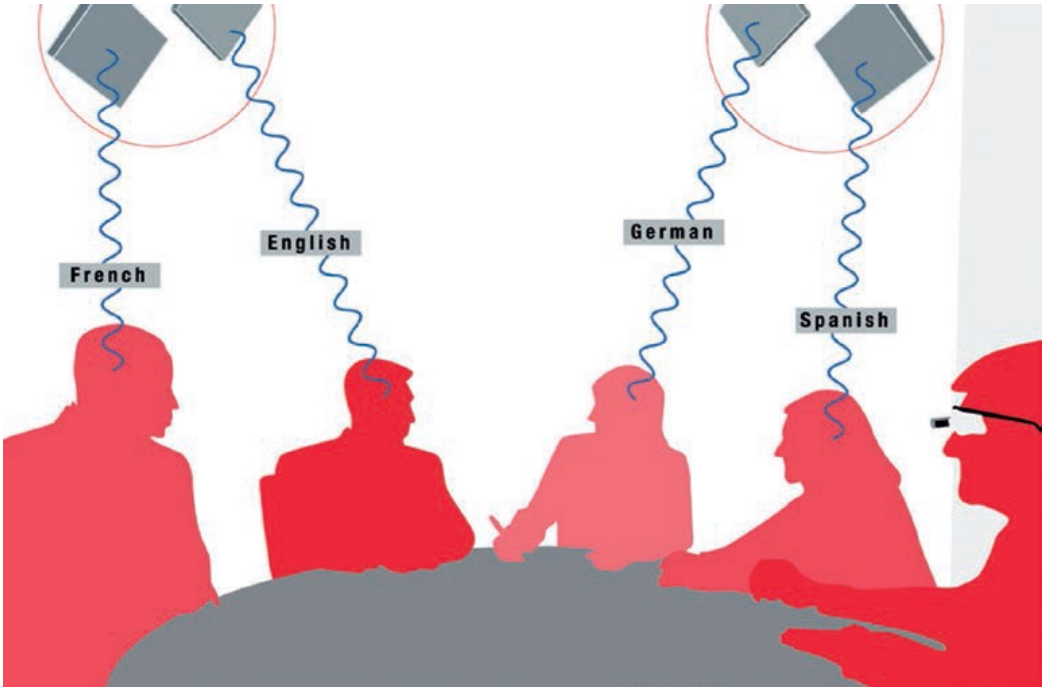

Figure 15: Individually adapted simultaneous translation without headset: With the help of directional microphones and heads-up display goggles.

Summary and Outlook

Modern language technologies are on the way of overcoming language barriers. A few years ago, this vision was considered to be unsolvable and appeared to be science fiction, as language barriers have separated mankind with a Babylonian confusion of languages for centuries now.

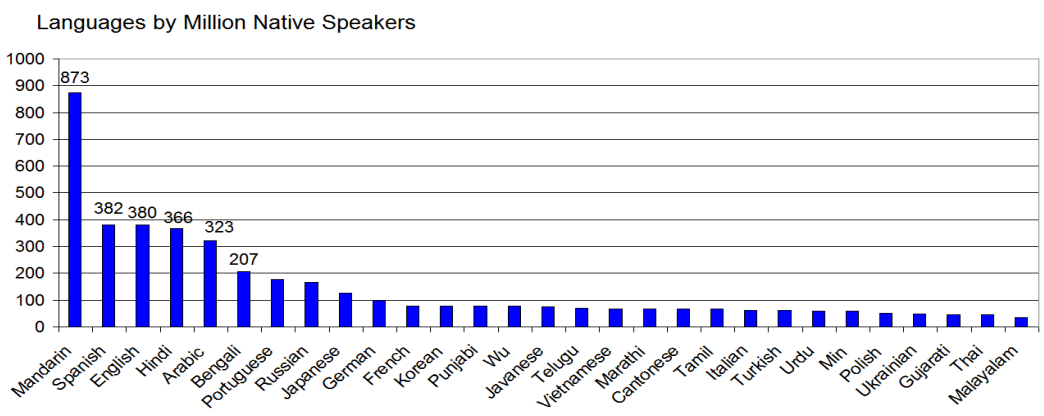

Figure 16: Use of languages in the world.

Much still remains to be done. Current translation technologies cover a few (~ 50) of the about 6,000 languages in the world. To cover all languages of the world, the technologies presented here still have to be "ported" to many other languages. To achieve this, construction of the systems has to become far more inexpensive and easier. Multilingual modeling [Bao et al., 2014], adaptation, language-independent models, better interaction with users, and self-learning systems will make this vision come true. Prognoses are good: The technologies described in this contribution can all be "learnt", which means that they can be trained automatically for every language using databases. Collection of sufficiently large data volumes is most important. The internet makes this easier and more realistic: Data can be collected remotely or they are produced indirectly by crowd sourcing, activities, games, or user feedbacks on a voluntary basis and at no cost. Today, modern translation systems are trained with more data (>> 1 GWords) than spoken by man on the average during lifetime (~ 0.5 GWords). This volume will further increase in the future.

Hence, chances are rather good for new algorithms and technologies bringing us closer together not only physically and enabling communication with every human being on our

planet. In this way, language barriers and understanding problems will be overcome in our generation already.

Acknowledgment

The present contribution was improved substantially by discussion with and reading and editorial support of Jan Niehues, Margit Rödder, Maria Schmidt, and Dorothea Schweizer. The author is grateful for their support and participation.

Literature

[...]