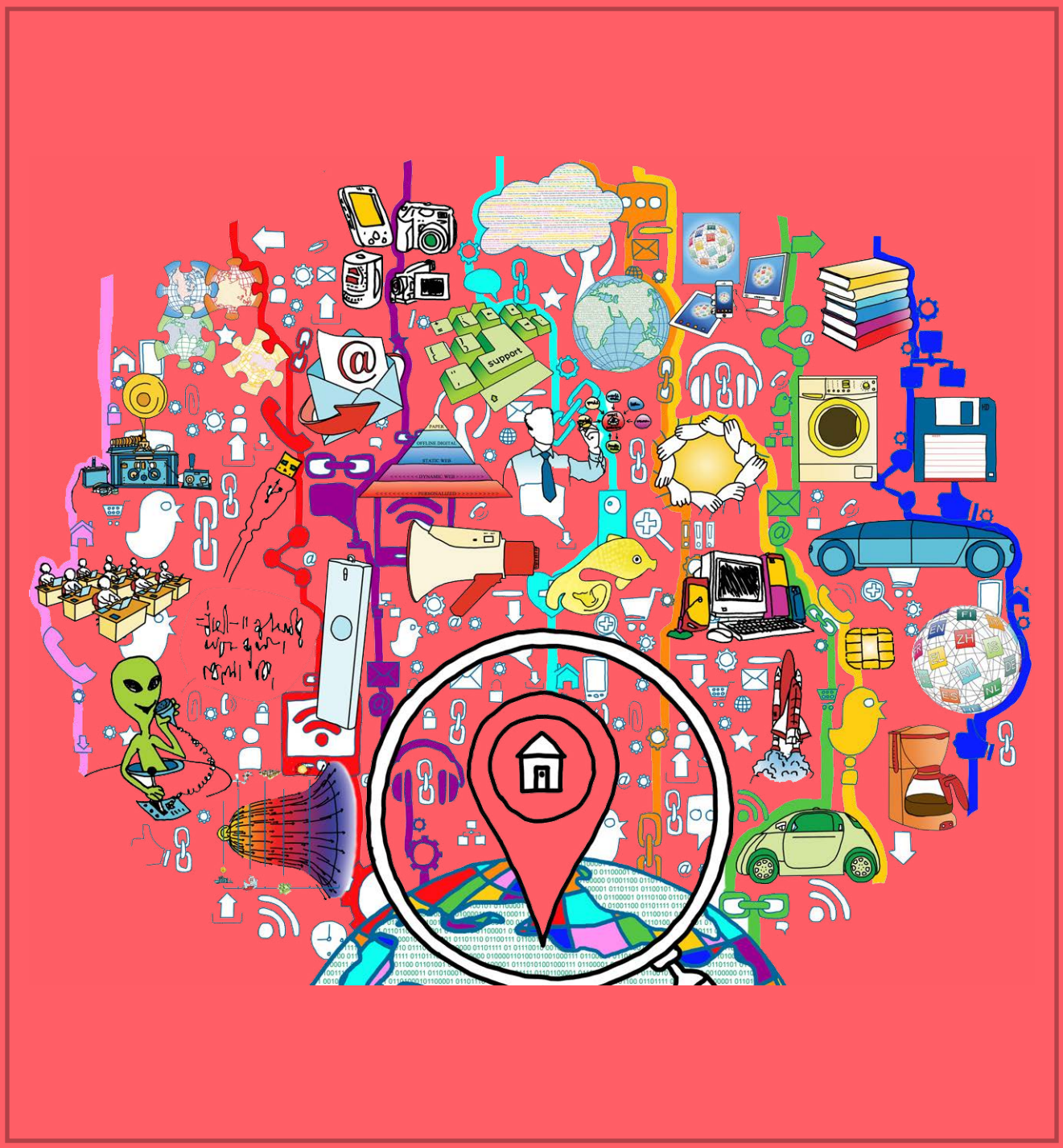
TAUS Speech-to-Speech Translation Technology Report

March 2017



TAUS
Enabling Better Translation

Authors: Mark Seligman, Alex Waibel, and Andrew Joscelyne
Contributor: Anne Stoker

Design: Anne-Maj van der Meer

Published by TAUS BV, De Rijp, The Netherlands

For further information, please email info@taus.net

Table of Contents

Introduction

The dream of **automatic speech-to-speech translation** (S2ST), like that of automated translation in general, goes back to the origins of computing in the 1950s. Portable speech translation devices have been variously imagined as *Star Trek*'s "universal translator" to negotiate extraterrestrial tongues, Douglas Adams' Babel Fish in the *Hitchhiker's Guide to the Galaxy*, and more. Over the past few decades, the concept has become an influential meme and a widely desired solution – not far behind the video phone (it's here!) and the flying car (any minute now).

Back on planet Earth, real-world S2ST applications have been tested locally over the past decade to help medical staff talk with other-language patients; to support military personnel in various theaters of war; to support humanitarian missions; and in general-purpose consumer products. A prominent recent project aims to build S2ST devices to enable cross-language communications at the 2020 Olympics in Tokyo, with many more projects and use cases in the offing. Automated speech translation has arrived: the tech's entry into widespread use has begun, and enterprises, app developers, and government agencies are alive to its potential.

More broadly, the recent spread of technologies for real-time communication – "smart" devices enabling on-the-spot exchanges using voice or text via smartphones – has helped promote the vision of natural communication on a globally connected planet: the ability to speak to someone (or to a robot/chatbot) in your language and be immediately understood in a foreign language. For many commentators and technology users, inspired by new models of deep learning, cognitive computing, and big data – and despite the inevitable doubts about translation quality – it seems only a question of time until S2ST becomes a trusted, and even required, communication support technology.

In view of this general interest in instant automatic speech translation services, TAUS believes that developers, enterprises, and the language technology supply community now need:

- a clear picture of the technological state-of-play in S2ST
- information on the history of this technology program
- an informed overview of the drivers and enablers in the field
- the near-term predictions of major and minor players concerning solutions and services, along with their assessments of weaknesses and threats

Accordingly, this TAUS report on S2ST provides an up-to-date account of the field's

technologies, approaches, companies, projects, and target use cases.

The report is part of an ongoing series (including the TAUS Translation Technology Landscape Report (2013 and 2016) and the TAUS Translation Data Landscape Report (2015)) providing state-of-the-art surveys of the relevant technologies, players, underlying vision, and market strengths and weaknesses. It doesn't predict market size or specific economic benefits, but does survey experimental business models.

Chapters follow on the Past, Present, and Future of speech-to-speech translation. The chapter on the Present contains interviews with 13 representative participants in the developing scene. An Appendix displays the results of a survey of potential users concerning anticipated uses of the technology.

A Note on Terminology

So far, there's no standardized way of talking about automatic speech-to-speech translation. Candidate terms include "speech translation" and "spoken (language) translation (SLT)," but these don't underscore the automaticity or underlying digital technology. "Automatic interpretation" (as inspired by human interpreting, e.g. in conferences) hasn't caught on, possibly because "interpretation" has other distracting meanings in English.

We'll use S2ST here for maximum clarity, but for variety will alternate with all of the above terms when the meaning is clear.

Past

This chapter of the TAUS report on S2ST recaps the history of the technology. Later chapters will survey the present and look toward the future.

The field of *speech* – as opposed to *text* – translation has an extensive history which deserves to be better known und understood. Text translation is already quite difficult, in view of the ambiguities of language; but attempts to automatically translate *spoken* rather than *written* language add the considerable difficulties of converting the spoken word into text (or into a semantic or other internal representation). Beyond the need to distinguish different meanings, systems also risk additional errors and ambiguity concerning what was actually said – due to noise, domain context, disfluency (errors, repetitions, false starts, etc.), dialog effects, and many more sources of uncertainty.

They must not only determine the appropriate meaning of "bank" – whether "financial institution," "river bank," or other; they also run the risk of misrecognizing the word itself, in the face of sloppy speech, absence of word boundaries, noise, and intrinsic acoustic confusability. "Did you go to the bank?" becomes /dɪd͡ʒəgowdəðəbæŋk/, and each segment may be misheard in various ways: /bæŋk/ ⇢ "bang"; /gowdəðə/ ⇢ "goat at a"; and so on. This extra layer of uncertainty can lead to utter confusion: when a misrecognized segment (e.g. "Far East" ⇢"forest") is translated into another language (becoming e.g. Spanish: "selva"), only consternation can result, since the confused translation bears neither semantic nor acoustic resemblance to the correct one.

Orientation: Speech Translation Issues

As orientation and preparation for our historical survey of the speech translation field, it will be helpful to review the issues confronting any speech translation system. We'll start by considering several dimensions of design choice, and then give separate attention to matters of human interface and multimodality.

Dimensions of Design Choice

Because of its dual difficulties – those of speech recognition and machine translation – the field has progressed in stages. At each stage, attempts have been made to reduce the complexity of the task along several dimensions: range (supported linguistic flexibility, supported topic or domain); speaking style (read vs. conversational); pacing (consecutive vs. simultaneous); speed and latency (real-time vs. delayed systems); microphone handling; architecture (embedded vs. server-based systems); sourcing (choice among providers of components); and more. Each system has necessarily accepted certain restrictions and

limitations in order to improve performance and achieve practical deployment.

Range (supported linguistic flexibility, supported topic or domain)
Restricted syntax, voice phrasebooks: The most straightforward restriction is to severely limit the range of sentences that can be accepted, thereby restricting the allowable syntax (grammar). A voice-based phrase book, for example, can accept only specific sentences (and perhaps near variants). This limitation does simplify recognition and translation by reducing the number of possible choices (in the jargon, the *perplexity*). Speech recognition need only pick one of the legal words or sentences, and translation requires no more than a table lookup or best-match operation. However, while these constraints improve performance and hence ease deployment, deviations from the allowable sentences will quickly lead to failure (though fuzzy matching can raise flexibility a bit). Thus voice-activated phrasebooks are effective in simple tasks like command-and-control, but can't handle free conversations, dialogs, speeches, etc.

Restricted-domain dialogs: Systems can limit the domain of a dialog rather than the range of specific sentences. Practical applications are those in which dialogs remain in a specific transactional domain and aim at a specific outcome, including registration desk and hotel reservation systems, facilities for scheduling or medical registration, and so on. Such systems impose fewer restrictions than phrasebooks, since users can in theory say anything they like ... *if* they remain within the supported topic or domain. And, unlike voice phrasebooks, restricted systems don't require users to remember allowable phrases or vocabularies – a requirement generally impractical for use cases involving untrained users, patients, or customers.

Domain restrictions simplify the work of developers, too: for both recognition and translation, we know the typical transactional patterns; can apply domain-dependent concepts and semantics; and can train appropriate models given large data and corpora from dialogues in that domain. Even so, limited-domain dialog systems are typically more difficult to engineer than those limited to phrasebooks, as they include varied expressions; greater disfluency and more hesitations ("I, er, I uhm, I would like to, er ... can I please, er .... Can I make a reservation, please?"); and generally less careful speech.

Open-domain speech: In open-domain systems we remove the domain restriction by permitting any topic of discussion. This freedom is important in applications like translation of broadcast news, lectures, speeches, seminars, and wide-ranging telephone calls. Developers of these applications confront unrestricted, and thus much larger, vocabularies and concept sets. (Consider, for example, special terms in academic lectures or speeches.) Moreover, open-domain use cases must often handle long monologues or continuous streams of speech, in which we don't know the beginnings and endings of sentences.

Speaking style (read vs. conversational speech): Among open-domain systems, another dimension of difficulty is the clarity of the speech – the degree to which pronunciation is well articulated on one hand, or careless and conversational on the other. The speech of a TV anchor, for example, is mostly read speech without hesitations or disfluencies. Given this clarity, it can be recognized with high accuracy, even when large vocabularies are in use. Lectures are harder: they aren't pre-formulated and some lecturers' delivery is halting and piecemeal. At the limit, spontaneous and conversational dialogs like meetings tend toward even more fragmentary and poorly articulated speech. (Mumbling is endemic.)

Pacing (consecutive vs. simultaneous): In *consecutive* speech translation, a speaker pauses after speaking to give the system (or human interpreter) a chance to produce the translation. In *simultaneous* interpretation, by contrast, recognition and translation are performed in parallel while the speaker keeps speaking. Consecutive translation is generally easier, since the system knows where the end of an utterance is. In addition, articulation is generally clearer, because speakers have time to formulate each utterance and can try to cooperate with the system: they can try to anticipate the output so as to be understood.

In simultaneous interpretation, the inverse is true: speakers are less aware of the system and less prone to cooperate.

Speed and latency (real-time vs. delayed systems): Difficulties may arise from a given task's speed requirements: latency (waiting time) may be intolerable beyond a certain threshold. For simultaneous speech interpretation, the system mustn't fall too far behind the speakers; and it may be desirable to produce a segment's translation as soon as possible after its pronunciation, perhaps before the end of the full utterance. Of course, low-latency interpretation is hard because it demands accurate rendering of the early segments before later segments become available to supply complete context. Fortunately, use cases differ in their demands: when an audience is following along *during* a lecture, parliamentary speech, or live news program, speech and low latency are indeed essential; but if the same discourses are audited *after the fact* for *post-hoc* viewing or browsing, there's no such need, and a system can use the entire discourse as context to produce the most accurate output.

Microphone handling: Speakers can sometimes use microphones reasonably close to them or attached, yielding relatively clear speech signals – e.g. in telephony, in lectures with headset or lapel microphones, and in mobile speech translators. Similarly, broadcast news utilizes studio-quality recording, quite amenable to today's recognition technology. However, performance rapidly degrades when speakers are far from their mics, or when there's *cross-talk* (overlap) among several speakers – as when table mics are used in meetings, or in recordings of free dialog captured "in the wild" with distant microphones, mobile phones, or cameras.

Architecture (mobile vs. server-based systems): Must speech translation technology run embedded on a mobile device, or is a network-based solution practical? Good performance is generally easier to engineer in networked implementations, because more extensive computing resources can be brought to bear, so that powerful speech translation capabilities can be delivered worldwide without heavy software downloads. Network-based solutions also enable collection of data from the field. On the other hand, in many speech translation applications, such solutions may be unacceptable – for example, when network-based processing is unavailable (e.g. for humanitarian assistance in remote areas); or too expensive (when roaming while traveling abroad); or insufficiently confidential or secure. For interpretation of lectures and speeches or for broadcast news, network-based solutions typically work well; by contrast, in applications for travel or for medical, humanitarian, military, or law-enforcement apps, embedded mobile technology is often preferable.

Sourcing (choice among providers of components): We've been discussing the implications of mobile vs. server-based architecture for system development and usage. Architecture choices also have organizational and business implications: in particular, where will the technology – the speech, translation, and other components – come from? Given the global character of the field, it has become possible for speech translation vendors to build applications without owning those components.

A vendor may for example build an interface that captures the voice utterance; sends it to an Internet language service (e.g. Nuance, Google, Microsoft, etc.) to perform speech recognition; sends the result to another service to perform machine translation; and finally sends it to a third service for speech synthesis. An embedded system might similarly be built up using licensed components. With either architecture, value might (or might not) be added via interface refinements, customization, combination of languages or platforms, etc. This systems integration approach lowers the barrier of entry for smaller developers, but creates a dependency upon the component providers which might become a liability if a use case requires facilities that the providers don't provide – new languages, specific vocabularies, and so on.

In the face of all these dimensions of difficulty and choice, speech translation solutions differ greatly. Each must match its use cases with the most appropriate technology. As a result, direct comparison between systems becomes difficult, and there can be no

simple answer to the question, "How well does speech translation work today?" In a given use case, depending on the relevant dimensions of difficulty, the answer can range from "Great! Easy problem, already solved!" all the way to "Not so good. Intractable problem, research ongoing." In response to the technical challenges posed by each dimension, speech translation as a field has progressed in stages, from simple voice-activated command-and-control systems to voice-activated phrasebooks; from domain-limited dialog translators to domain-unlimited speech translators; from demo systems to fully deployed networked services or mobile, embedded, and general-purpose dialog translators; and from consecutive to simultaneous interpreters.

Human Factors and Interfaces

Speech Translation is after all an aid to communication among humans, so of course the human interface is essential. Ideally, we'd want to simply hear and understand conversation partners as if they were speaking our language and the translation program weren't there: the task of the interface is to make the language barrier as transparent as possible. We want maximum speed and minimum interference on one hand, while maintaining maximum accuracy and naturalness on the other. These are of course competing goals. Good interface solutions can help to balance them; but no perfect solutions are to be expected in the near term, since even human interpreters normally spend considerable time in clarification dialogues.

As long as perfect accuracy remains elusive, efficient error recovery mechanisms will remain desirable. The first step is to enable users to recognize errors, both in speech recognition and in translation. To correct errors once found, mechanisms for correction, and then for adaptation and improvement, are needed.

Speech recognition errors can be recognized – by literate users – if speech recognition results are displayed on a device screen. For illiterate users, or to enable eyes-free use, text-to-speech playback of ASR results could be used (but has been used only rarely to date). To correct ASR mistakes, some systems may enable users to type or handwrite the erroneous word. Facilities might instead be provided for voice-driven correction (though these, too, have been used only rarely to date). The entire input might be repeated – but then the same errors might recur, or new ones might erupt. Finally, multimodal resolutions can be supported, for instance involving manual selection of an error in a graphic interface followed by voiced correction. (More on multimodal systems just below.)

In any case, if a segment or utterance can be corrected, it can be passed to machine translation (at least in systems whose ASR and MT components are clearly separate). Then recognition and correction of translation results may be facilitated.

Several SLT systems aid recognition of machine translation errors by providing indications of the system's confidence: low confidence flags potential problems. In a similar spirit, other systems supply back-translations, so that users can determine whether the input is still understandable after its round trip through the output language. (However, back-translation can introduce additional errors, thus yielding misleading – and often comical – results.

Some systems have minimized such extraneous mistakes by generating back-translations directly from language-neutral semantic representations. And one system has developed techniques for enhancing the accuracy of back-translations in systems lacking such interlingua representations: when generating the reverse translation, the MT engine is constrained to reuse the semantic elements used in the forward translation.)

User-friendly facilities for real-time *correction* of translation errors are challenging to design. They may include tools for lexical disambiguation, or choice among available meanings for ambiguous expressions. In one system, for example, if "This is a cool program!" mistakenly yields "This is a chilly program" as shown by back-translation, the user can interactively select the desired meaning for "cool" by reference to synonym cues – choosing e.g. "awesome, great" in preference to "chilly, nippy." Such interaction can be distracting,

so means can be provided for indicating the desired degree of interactivity.

Some systems have experimented with robot avatars designed to play the role of mediating interpreters who (which?) could interact with users to attempt resolution of translation mistakes. In one such system, intervention was ultimately judged too distracting, and a design has been substituted in which users recognize errors by reference to a running transcript of the conversation, augmented by their own perception of misunderstandings: when these occur, rephrasing and questioning are encouraged.

Whether in ASR or MT, errors are annoying, but repeated errors are *damn* annoying. Ideally, systems should learn from their mistakes, so that errors diminish over time and use. If machine learning is available, it should take advantage of any corrections that users supply. (Dynamic updating of statistical models is an active research area.) Alternatively, interactive update mechanisms can be furnished.

One more interface issue involves frequent recurrence of a given utterance, e.g. "What is your age?" Verbal repetition can quickly become inefficient, especially if mistakes recur. Accordingly, translation memory (TM) can be supplied in various forms: most simply, a system can record translations for later reuse.

Multimodal Translators

The ultimate purpose of an SLT system is to provide flexible and natural cross-language communication. Clearly, this communication may involve more than text and speech. A wide range of modalities for both input and output may come into play.

On the input side, systems can translate not only speech but text messages, posts[1], images of road-signs and documents [Yang et al., 2001a, 2001b; Zhang et al., 2002a, 2002b; Waibel, 2002; Gao et al., 2004][2] ... even silent speech by way of muscle movement of the articulators, as measured through electromyographic sensors! [Maier-Hein et al., 2005][3]. Going forward, multimodal input will be needed to better capture and convey human elements of communication: emotions, gestures, and facial expressions will help to transmit speakers' intent in the context of culture, relationships, setting, and social status. Research in these areas is ongoing.

Multimodal *output* choices will likewise vary per situation. In lectures, for example, audible speech output from multiple sources would be disruptive, so preferable delivery modes may involve headphones, targeted audio speakers that project only within a narrow cone[4], etc. Text may be preferred to spoken output, or may be added to it – on personal devices, in glasses or goggles for heads-up display.

Chronology and Milestones

Having gained some perspective on the issues facing speech translation systems – the design choices and considerations of human interface and multimodality – we can now begin our historical survey.

Fictional systems had already put a twinkle in many eyes, but the earliest actual demonstration seems to have been in 1983, when the Japanese company NEC presented a system at that year's ITU Telecom World[5]. This demonstration system employed domain-limited phrasebooks, and thus gave an incomplete preview of coming attractions, but it did illustrate the vision and feasibility of automatically interpreting speech.

Further progress would await the maturation of the main components of any speech translation system – speech recognition, machine translation, speech synthesis, and a viable infrastructure. Fully functional continuous speech recognition for large vocabularies began to emerge only at the end of the '80s. At that point, text-based machine translation was still an unsolved – and, in the view of many, unsolvable – problem: it was seriously

1 E.g. in Facebook: https://www.facebook.com/help/509936952489634?helpref=faq_content

2 E.g. in Google: https://support.google.com/translate/answer/6142483?hl=en

3 See https://www.youtube.com/watch?v=aMPNjMVl-r8A

4 E.g. in the work of Jörg Müller and others: https://www.newscientist.com/article/mg22129544-100-beams-of-sound-immerse-you-in-music-others-cant-hear/.

5 See for example http://link.springer.com/article/10.1007/s40012-013-0014-4 or https://itunews.itu.int/En/2867-TELECOM-83BRTelecommunications-for-all.note.aspx.

attempted again only after a multi-decade hiatus in the late '80s and early '90s. Meanwhile, unrestricted speech synthesis was just appearing [Allen et al., 1979]. Also emerging was a medium for transmission: several companies – Uni-verse, Amikai, CompuServe, GlobalLink, and others – attempted the first chat-based text translation systems, designed for real-time use but lacking speech elements[6].

By the early '90s, speech translation as a vision had generated sufficient excitement that research in the space was funded at the national level and began in earnest. In Japan, the Advanced Telecommunications Research (ATR) Institute International opened officially in April, 1989, with one of its four labs dedicated to Interpreting Telephony. A consortium underwritten by the Japanese government brought together investment and participation from a range of Japanese communication firms: NTT, KDD, NEC, and others[7].

Researchers from all over the world joined the effort, and collaborative research with leading international labs was initiated. The Consortium for Speech Translation Advanced Research (C-STAR) was established in 1992 by ATR (initially under the direction of Akira Kurematsu), Carnegie Mellon University in Pittsburgh (CMU) and the Karlsruhe Institute of Technology (KIT) in Germany (coordinated by Alexander Waibel), and Siemens Corporation. In January, 1993, the same group mounted a major demo linking these efforts as the culmination of an International Joint Experiment on Interpreting Telephony. It was widely reported – by *CNN*, the *New York Times*, *Business Week*, and many other news sources – as the first international demonstration of spoken language translation, showing voice-to-voice rendering via dedicated long-distance video hook-ups for English⇨Japanese, English⇨German, and Japanese⇨German. The immediate goal was again a proof of concept – a demonstration that the dream of breaking language barriers through technology might one day be realized.

At the Japanese end, the speech translation system was named ASURA, for a many-faced Buddhist deity [Morimoto et al., 1993]. ASURA's speech recognition, based on hidden Markov models, yielded a ten-best list for the input utterance as a whole, from which the speaker chose the best candidate. The system's translation component was entirely rule-based: analysis, transfer, and generation exploited unification in the style of Martin Kay [Shieber, 2003].

Analysis results were intended to represent relatively deep semantics, since they were intended to mediate between quite different languages; nevertheless, many surface-language elements were retained [Seligman et al., 1993]. The combined ASURA system was potentially powerful, but extremely brittle in that its hand-built lexicons were narrowly restricted to the selected domain, conference registration. The system was also slow, due to the hardware limitations of the time and the computational demands of unification; as a result, it was retired soon after the demo, though Moore's Law might soon have come to its rescue.

In the demo, the speech translation system for German and English was the first in the US and Europe, coincidentally named for another two-faced god, JANUS [Waibel et al., 1991, 1997][8]. The system was also one of the first to use neural networks for its speech processing. Analysis for translation was performed in terms of Semantic Dialog Units, roughly corresponding to speech acts. Two parsers were used, the first designed for accuracy and using a more restrictive syntax, and the second intended for robustness and based on semantic parsing. For both parsers, output was transformed into a common semantic exchange format. This representation yielded two advantages: first, domain semantics could be enforced, so that all sentences, even if stuttered or otherwise disfluent, would be mapped onto well-defined

6 Translating chat systems have survived to the present, spurred by the recent explosion of texting. San Diego firm Ortsbo, for example, is primarily a chat aggregator, supplying a bridge among many different texting platforms; but it also enables multilingual translation of the various streams and – looking ahead to speech translation – has purchased an interest in Lexifone, a speech translation company we'll encounter below.

7 This consortium followed upon another ambitious government-sponsored R&D effort in the 1980s: the Fifth Generation project, which aimed to build computers optimized to run the Prolog computer language as a path toward artificial intelligence.

8 Its name, beyond the classical reference, also served as a tongue-in-cheek acronym for Just Another Neural Understanding System.

semantic concepts to support generation of understandable utterances in the output language; and second, the semantic representation could also serve as a language-neutral *interlingua*, so that additional languages could be added relatively easily, and translation could be facilitated along all possible translation paths.

Research ambitions at the participating sites extended well beyond simple concatenation of the major speech translation components – speech recognition, machine translation, and text-to-speech. A wide variety of other studies tackled example-based translation, topic tracking, discourse analysis, prosody, spontaneous speech features, neural network-based and statistical system architectures, and other aspects of an idealized translation system. ATR, in cooperation with CMU and Karlsruhe, also amassed a substantial corpus of close transcriptions of simulated conversations to serve as reference and training for automated components.

As the 1993 demo was taking shape, the parties also expanded the international C-STAR cooperation into its second phase. To the original US, German, and Japanese research groups (Carnegie Melon University, Karlsruhe Institute of Technology, Advanced Telecommunications Research International) were added organizations from France (GETA-CLIPS, University Joseph Fourier); Korea (Electronics Telecommunications Research Institute); China (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences); and Italy (ITC-IRST, Centro per la Ricerca Scientifica e Tecnologica). Over the following decade, plenary meetings were held annually to compare and evaluate developing speech translation systems. Locations included Pittsburgh, Grenoble, Taejon (Korea), Guilin (China), Trento (Italy), Munich, and Geneva[9].

To facilitate comparison among C-STAR systems, all employed the same underlying representation for the meanings of utterances. The representation aimed for language neutrality – that is, it was an interlingua, in this case abbreviated IF for Interchange Format [Levin et al., 1998]. Mediation between this pivot and the input and output surface language was handled by each system in its own way.

The use of a common interlingua, however, also had two drawbacks. First, it was necessary to develop and maintain this representation and the parsers mapping into it – at the time, manually. Second, the representation was domain-limited, so that the resulting systems could operate only in the chosen domains (hotel reservation, travel planning, etc.). Hand-coded parsers were gradually replaced by parsers trainable via machine learning, but the limitation to specific domains remained, yielding systems appropriate only for tightly goal-driven transactional dialogs.

The need became apparent for extension to domain-independent tasks. Accordingly, as an alternate approach, the consortium also became the context for the first practical demonstration, in Grenoble in 1998, of unrestricted or open-ended speech translation. Under the auspices of the French team, the demo was built by adding local speech input and output elements to a server-based chat translation system for several European languages created by CompuServe under the management of Mary Flanagan [Seligman, 2000]. The resulting SLT system enabled users to interactively correct recognition errors and incorporated large-vocabulary commercial-grade speech recognition and translation components.[10]

C-STAR also became the venue for early experiments with novel communication channels. In Guilin, China, in 2002, the Korean C-STAR group demonstrated the first speech translation via a cellular phone connection. The demo

9 And quite enjoyable they were. Someone suggested renaming the association as Consortium for Sightseeing, Travel, and Restaurants.

10 One previous demonstration of open-ended spoken language translation was reported formally only by a third party [Seligman, 1996]. This was the Information Transcript Project, part of an artistic exposition, the Biennale d'Art Contemporain in Lyon, France, in the winter of 1995-1996. Francophone viewers at the exposition could say whatever they chose into a microphone one word at a time, and counterparts at MIT in Cambridge, Massachusetts could hear spoken translations into English – and vice versa. CU-SeeMe let speakers see and hear each other. For speech recognition, IBM Voice Type was used. Translation was carried out by GlobalLink translation software and translated text was transported across the Atlantic by FTP (file transfer protocol). Apple's Macintosh text-to-speech completed the speech-to-speech cycle.

foreshadowed the first two SLT products for telephony, which entered the Japanese market four and five years later: NEC's mobile device for Japanese-English (2006) and the Shabete Honyaku service from ATR-Trek (2007)[11].

An early boxed product for speech translation on PCs used similar component technologies. This was *talk & translate* for German<->English, produced by Linguatec (later Lingenio) in 1998. To the company's own bidirectional translation software were added ViaVoice from IBM and its associated text-to-speech. The product suffered from the difficulty of individual speech registration – a twenty-minute training session was needed for the speaker-dependent software of the time – and failed to find a market in German business, where English competence was already widespread at the managerial level[12].

Germany also became the scene of a major government-supported speech translation endeavor during the '90s – the Verbmobil project, headed by Wolfgang Wahlster and Alex Waibel and sponsored by the German Federal Ministry of Research and Technology from 1993[13]. Verbmobil adopted many of the techniques and subsystems already explored under C-STAR, but conducted extensive further research throughout Germany in an attempt to advance the state of the art. As in the C-STAR consortium, studies were undertaken of discourse, topic tracking, prosody, incremental text generation, and numerous other aspects of a voice-to-voice translation system assumed to require many cooperating knowledge sources.

Once again, however, integration proved difficult: the combination of many knowledge sources became unwieldy and hard to maintain. One element of the research program, however, did prove seminal for later systems: the use of statistical machine translation (SMT) for speech translation. Originally proposed for text translation at IBM [Brown et al., 1993], the approach was championed by Verbmobil researchers at Karlsruhe Institute of Technology (KIT), Carnegie Mellon University (CMU) [Waibel, 1996], and the Rheinisch-Westfälische Technische Hochschule (RWTH) [Och and Ney, 2002], who then further developed SMT to create the first statistical *speech* translators. Statistical techniques enabled the training of the associated MT systems from parallel data without careful labeling or tree-banking of language resources. Although the significance of the approach was not immediately recognized – Wolfgang Wahlster, for instance, argued that "While statistical translation ... produces quick and dirty results ... semantic transfer ... produces higher quality translations ..." [Wahlster, 2000, page 16] – SMT often yielded superior performance and better translation quality than rule-based methods, due partly to the consistency of its learning. The approach went on to become dominant within C-STAR and other programs.

Other noteworthy projects of the time:

- A prototype speech translation system developed by SRI International and Swedish Telecom for English-Swedish in the air travel domain [Alshawi et al., 1992];
- The VEST system (Voice English/Spanish Translator) built by AT&T Bell Laboratories and Telefonica Investigacion y Desarrollo for restricted domains [Roe et al., 1992] (using finite state transducers to restrict language processing in domain and syntax);
- KT-STS, a prototype Japanese-to-Korean SLT system created in 1995 by KDD in cooperation with Korea Telecom (KT) and the Electronics and Telecommunications Research Institute (ETRI) in Korea, also for limited domains[14].

Approaching the present, we come to twin watersheds which together shape the current era of speech translation: the advent of big data and of the app market.

As big data grew ever bigger, Google Translate took off in 2006-2007 as a translator for text (initially in Web pages), switching from earlier rule-based translation systems to statistical MT under the leadership of Franz Och (who had developed early SMT under Verbmobil and at ISI, the Information Science Institute). In this effort, general-purpose, open-domain MT made great strides: machine translation

11 See the next chapter (Present) concerning current phone-based systems by SpeechTrans and Lexifone.

12 Subsequent products have fared better, however.

13 See http://verbmobil.dfki.de/overview-us.html.

14 See http://tinyurl.com/gwf286s.

became available to every Internet user, and this accomplishment unarguably marked the most dramatic and influential milestone in worldwide public use of the technology. Equally significant was SMT's broad adoption as the method of choice for MT, as Google took full advantage of its massive databanks. (Research Director Peter Norvig is often misquoted as saying, "We don't have better learning algorithms than everybody else, just more data." He was actually quoting Michele Banko and Eric Brill and espousing a more nuanced view[15].)

Och oversaw a massive expansion of words translated and language pairs served. (The service now bridges more than one hundred languages and counting, and translates 100 billion words per day.) Although in the previous decade machines had translated only a negligible percentage of texts, they were soon to generate 99% of the world's translations. True, the results were often inferior to human translations; but they were clearly improving, often sufficiently understandable, readily available and ... free of cost (!) for everyone's use via the Internet. Translation of speech, however, was not yet attempted at Google. (At IWSLT'08, Och argued at a panel discussion against its feasibility and readiness and usefulness to Google as a Web company.)

Then came the transition to Mobile. As mobile phones became smartphones and the mobile app market took shape, sufficient processing punch was finally packed into portable or wearable devices to enable creation of fully mobile and embedded speech translators. With the advent of the iPhone 3G, advanced speech and machine translation technology could fit on a phone. A newly enabled system could exploit advanced machine learning and include sufficiently large vocabularies to cover arbitrary traveler needs.

In 2009, Mobile Technologies, LLC, a start-up company founded by Alex Waibel and his team in 2001, launched Jibbigo [Eck et al., 2010], the first speech translator to run entirely without network assistance on iPhone and Android smartphones. The product featured a 40,000 word vocabulary and produced voice output from voice input faster than a chat message could be typed. While it was designed with travelers or healthcare workers in mind, it was domain-independent and thus served as a general dialog translator.

The first Jibbigo app provided open-domain English⇨Spanish speech translation and offered a number of user interface features for rapid error recovery and rapid expansion of vocabularies in the field: for instance, back-translations – secondary translations from the output (target) language back into the input (source) language – helped users to judge translation accuracy. The app incorporated customization features as well: users could enter proper names missing from a system's ASR vocabulary (like "München" in English and "Hugh" in German), thereby automatically converting and loading associated elements (dictionaries, models of word sequence, etc. in ASR, MT, and text-to-speech) without requiring linguistic expertise [Waibel and Lane, 2012a, 2012b, 2015]; and system extensions for humanitarian missions featured user-definable phrasebooks that could be translated by machine and then played back at will [Waibel and Lane, 2015].

Because Jibbigo incorporated machine learning technology, it could rapidly add 15 languages in the following two years. And because of its network independence, it could be used extensively where networks were unavailable, for example by travelers and healthcare workers in humanitarian missions (2007-2013) [Waibel et al., 2016]. Apple commercials for iPhones featured it extensively. A free network-based version with chat capabilities was also provided. The company was subsequently acquired by Facebook and formed the basis of the Facebook Language Technology Group in 2013.

Soon after Jibbigo's appearance, Google entered the SLT field with a network-based approach to mobile speech translation. Conversation Mode was demonstrated in 2010 and released in an alpha version for English ⇨Spanish in early 2011. By October of that year, the service expanded to 14 languages. Microsoft, too, launched speech translation apps.

15 See https://www.quora.com/Does-Google-not-have-better-algorithms-but-only-more-data.

In the 2010s also, in systems addressing limited domains, continuing movement could be seen from research programs – constrained by the limitations of the technology – toward those designed for commercial use in vertical markets, purposely focused for practicality within a use case.

Healthcare presents a particularly challenging vertical. An early challenger was the S-MINDS system by Sehda, Inc. (later Fluential) [Ehsani et al., 2008]. At its center was an extensive set of fixed and pre-translated phrases; and the task of speech recognition was to match the appropriate one so as to enable pronunciation of its translation via text-to-speech. A proprietary facility yielded the best available fuzzy match when no precise match was found. In this respect, the system represented further development of SLT systems like the earlier Phraselator[16], a ruggedized handheld device likewise offering translation of fixed phrases only, provided in large quantities to the US military for use in the first Gulf War and later in various military, law-enforcement, and humanitarian operations. Later versions of the Phraselator added licensed Jibbigo technology to provide more flexible speech input.

Other phrase-based systems designed for specific use cases included Sony's TalkMan[17] – a system sporting an animated bird as a mascot – and several voice-based phrasebook translators on dedicated portable devices from Ectaco, Inc[18].

Converser for Healthcare 3.0, a prototype by Spoken Translation, Inc., was pilot tested at Kaiser Permanente's Medical Center in San Francisco in 2011 [Seligman and Dillinger, 2011, 2015]. Converser provided full speech translation for English⇨Spanish. To overcome a perceived reliability gap in demanding areas like healthcare, business, emergency response, military and intelligence, etc., facilities for verification and correction were integrated: as in Jibbigo and MASTOR [Gao et al., 2006], users received feedback concerning translation accuracy in the form of back-translation; but Converser also applied semantic controls to avoid typical back-translation errors in MT engines lacking interlingua representation. If errors due to lexical ambiguity were found, users could interactively correct them using synonyms or other cues. To customize SLT for particular use cases, the system also included pre-translated frequent phrases. These could be browsed by category or subcategory (e.g. **Pickup** or **Consultation** within the **Pharmacy** category) or discovered via keyword search, and were integrated with full MT.

In 2004, DARPA launched a number of research programs in the United States to develop speech translators that government officers could use to communicate with local populations in areas of deployment for law enforcement and disaster relief, at checkpoints, etc. (Concerning Project DIPLOMAT, see [Frederking et al., 2000]; for Project BABYLON, see [Waibel et al., 2003]; for Project TRANSTAC, see [Frandsen et al., 2008].) In parallel, a large program was also launched to develop speech translation technology for use in translating and summarizing broadcast news. (Concerning Project GALE, see [Cohen, 2007] and [Olive et al., 2011].) Both programs were further advanced in a combined program, Project BOLT[19].

Initial efforts in DARPA programs (e.g. in DIPLOMAT) had developed only voice-based phrasebooks for checkpoints, but ongoing DARPA research programs BABYLON and TRANSTAC advanced to development of flexible dialog (in DARPA parlance, "two-way") speech translators. Several players – BBN, IBM, SRI, and CMU – participated. IBM's MASTOR system [Gao et al., 2006] incorporated full machine translation and attempted to train the associated parsers from tree-banks rather than build them by hand.

MASTOR's back-translation provided feedback to users on translation quality, making good use of an interlingua-based semantic representation. (The interlingua was also found helpful in compensating for sparseness of training data.) BBN, SRI, and CMU developed similar systems on laptops, while CMU also implemented on (pre-smartphone!) mobile devices of the day.

16 See https://en.wikipedia.org/wiki/Phraselator.
17 See https://en.wikipedia.org/wiki/Talkman.
18 See https://en.wikipedia.org/wiki/Ectaco.
19 Concerning Project BOLT, see http://www.darpa.mil/program/broad-operational-language-translation and https://www.sri.com/work/projects/broad-operational-language-technology-bolt-program.

Interestingly, systems developed and evaluated under Program BOLT demonstrated the feasibility of error correction via spoken disambiguation dialogs [Kumar, 2015]. However, voice fixes were found less efficient than typed or cross-modal error repair when available, thereby confirming conclusions drawn earlier for speech-based interfaces in general [Suhm et al., 1996a and 1996b].

Program GALE, perhaps the largest speech translation effort ever in the US, focused upon translation of broadcast news from Chinese and Arabic into English using statistical core technologies (ASR, MT, and summarization). The effort produced dramatic improvement in this tech, usable for browsing and monitoring of foreign media sources [Cohen, 2007; Olive et al., 2011].

Serious R&D for speech translation has continued worldwide, both with and without government sponsorship. Some notable efforts have included the following (with apologies to those not listed):

- Raytheon BBN Technologies [Stallard et al., 2011]
- IBM [Zhou et al., 2013]
- Nara Institute of Science and Technology (NAIST) [Shimizu et al., 2013]
- Toshiba[20]
- VOXTEC[21]
- Japan Global Communication Project, especially the National Institute of Information and Communications Technology (NICT)[22]

... which brings us to the present.

20 See http://www.toshiba.co.jp/about/press/2015_10/pr2901.htm.

21 See http://www.voxtec.com/wp-content/uploads/2015/03/Phraselator-P2-Product-Sheet.pdf.

22 See e.g. http://www.japantimes.co.jp/news/2015/03/31/reference/translation-tech-gets-olympic-push/#.WLtEfvnyu70 and https://www.taus.net/think-tank/articles/japan-s-translation-industry-is-feeling-very-olympic-today.

Present

Having traversed the history of automatic speech-to-speech translation (S2ST), we arrive at the somewhat paradoxical present.

On one hand, the technology has finally emerged from science fiction, research, and forecasts. It has finally become real, and is really in use by many. On the other hand, some knowledgeable parties still view the tech as not yet ready for prime time. Gartner, in particular, shows speech translation as at the Peak of Inflated Expectations, in spite of dramatic recent progress and despite current deployment in actual products and services.

Two factors seem to be at play in this caution and skepticism. First, large profits have remained difficult to identify. Second, current usage remains largely in the consumer sphere: penetration remains for the future in vertical markets like healthcare, business, police, emergency response, military, language learning, etc.

Both of these shortfalls seem to have the same origin: that the current era of speech translation is the age of the giants. The most dramatic developments and the greatest expenditure are now taking place at the huge computation/communication corporations. These have until now viewed SLT not as a profit center but as a feature for attracting users into their orbits — as honey to attract bees into their hives. S2ST has been included as added value for existing company services rather than as stand-alone technology.

The result has been, on one hand, stunning progress in technology and services; and, on the other, galloping commoditization. Consumers already expect S2ST to be free, at least at the present level of accuracy, convenience, and specialization. This expectation has created a challenging climate for companies depending on profit, despite continuing universal expectation that worldwide demand for truly usable speech translation will — sometime soon!— yield a colossal, and colossally profitable, world market.

Accordingly, this chapter on the present state of SLT will provide not a report on an established market, but rather a representative survey of the present activities and strivings of the large and small. We'll see incredible triumphs side by side with great expectations yet to be realized.

For this purpose, we've conducted a series of interviews, including both corporate representatives, large and small, and academic researchers. The participants, questions, and summarized results are below.

First though, as orientation, here's the briefest of snapshots – a selfie, if you like – of selected technical accomplishment at the current state of the art.

Google Translate mobile app:

- *Speed*: Barring network delays, speech recognition and translation proceed and visibly update while you speak: no need to wait till you finish. When you do indicate completion by pausing long enough – about a half second – the pronunciation of the translation begins instantly.
- *Automatic language recognition*: Manually switching languages is unnecessary: the application recognizes the language spoken – even by a single speaker – and automatically begins the appropriate speech recognition and translation cycle. End-of-speech recognition, too, is automatic, as just explained. As a result, once the mic is manually switched on in automatic-switching mode, the conversation can proceed back and forth hands-free until manual switch-off. (Problems will arise if speakers overlap, however.)
- *Noise cancellation*: Speech recognition on an iPhone works well in quite noisy environments – inside a busy store, for instance.
- *Offline capability*: Since speakers, and especially travelers, will often need speech translation when disconnected from the Internet, Google has added to its app the option to download a given language pair onto a smartphone for offline use.
- *Dynamic optical character recognition*: While this capability plays no part in speech translation per se, it now complements speech and text translation as an integral element of Google's translation suite. It enables the app to recognize and translate signs and other written material from images (photos or videos), rendering the translation text within the image and replacing the source text as viewed through the smartphone's camera viewer. The technology extends considerable previous research in optical character recognition (OCR), and builds on work by WordLens, a startup acquired by Google in 2014 that had performed the replacement trick for individual words. The current version handles entire segments and dynamically maintains the positioning of the translation when the camera and source text move.

Skype Translator (powered by Microsoft):

- *Telepresence*: Microsoft and its Skype subsidiary weren't the first to offer speech translation in the context of video chat: as one example, by the time Skype Translator launched, Hewlett-Packard had for more than two years already been offering a solution in its bundled MyRoom application, powered by systems integrator SpeechTrans, Inc. And speech translation over phone networks – but lacking video or chat elements – had been inaugurated experimentally through the C-STAR consortium and commercially through two Japanese efforts (as mentioned in the previous chapter). But the launch of Skype Translator had great significance because of its larger user base and consequent visibility – it exploits the world's largest telephone network – and in view of several interface refinements.
- *Spontaneous speech*: The Microsoft translation API contains a dedicated component, TrueText, to "clean up" elements of spontaneous speech – hesitation syllables, errors, repetitions – long recognized as problematic when delivered to an SLT system's translation engine. The component's goal, following a long but heretofore unfulfilled research tradition, is to translate not what you said, stutters and all, but what you meant to say.
- *Overlapping voice*: Borrowing from news broadcasts, the system begins pronunciation of its translation while the original speech is still in progress. The volume of the original is lowered so as to background it. The aim of this "ducking" is to encourage more fluid turn-taking. The hope is to make the technology disappear, so that the conversation feels maximally normal to the participants.

Interpreting Services InterACT (Waibel, KIT/CMU):

- *Simultaneous Interpreting Services*: Following its release of consecutive speech translators – including Jibbigo, the first network-free mobile SLT application – the team at InterACT (Waibel et al.) pioneered real-time and simultaneous automatic

interpreting of lectures. The technology was first demonstrated at a press conference in October, 2005 [Fügen et al., 2007; Waibel et al., 2013]. It was later deployed in 2012 as a Lecture Interpretation Service for German and now operates in several lecture halls of the Karlsruhe Institute of Technology (KIT). Target users are foreign students and the hearing-impaired.

- *Continuous online interpretation streaming*: During a lecture, speech is streamed over WiFi to KIT servers that process subscribed lectures. Speech recognition and Translation is performed in real time, and output is displayed via standard Web pages accessible to students.
- *Off-line browsing*: Transcripts are offered offline for students' use after class. Students can search, browse, or play segments of interest along with the transcript, its translation, and associated slides.
- *Speed*: The Lecture Translator operates at very low latency (time lag). Transcriptions of the lecturer's speech are displayed instantaneously on students' devices as subtitles, and translations appear incrementally with a delay of only a few words, often before the speaker finishes a sentence.
- R*eadability*: To turn a continuous lecture into readable text, the system removes disfluencies (stutters, false-starts, hesitations, laughter, etc.), and automatically inserts punctuation, capitalization, and paragraphs. (Speakers needn't pronounce commands like "Comma," "Cap That," or "New Paragraph.") Spoken formulas are transformed into text where appropriate ("Ef of Ex" ⇨ *f*(x)). Special terms are added to ASR and MT dictionaries from background material and slides.
- *Multimodality*: Beta versions include translation of slides; insertion of Web links giving access to study materials; emoticons; and crowd-editing. These versions also support alternative output options: speech synthesis, targeted audio speakers instead of headphones, or goggles with heads-up displays.
- *European Parliament*: Variants and subcomponents are being tested at the European Parliament to support human interpreters. (A Web-based app automatically generates terminology lists and translations on demand.) The system tracks numbers and names – difficult for humans to remember while interpreting. An "interpreter's cruise control" has been successfully tested for handling repetitive (and boring) session segments like voting.

Interviews

Now on to the aforementioned interviews. The interviewees, in alphabetical order by institution:

Interviewee	Date (in 2016)
Alibaba (Eric Liu)	1 July
Chinese Academy of Sciences (Chengqing Zong)	4 August
CMU/Karlsruhe Institute of Technology (Alex Waibel)	12 August
EML (Siegfried 'Jimmy' Kunzmann)	20 July
IBM/Microsoft ASG (Yuqing Gao)	1 August
Lexifone (Ike Sagie)	14 July
Logbar (Takuro Yoshida)	18 August
Microsoft/Skype (Chris Wendt)	5 July
NICT (Eiichiro Sumita)	28 July
Speechlogger (Ronen Rabinovici)	22 July
SpeechTrans (John Frei and Yan Auerbach)	12 July
Spoken Translation, Inc. (Mark Seligman)	8 August
Translate Your World (Sue Reager)	12 July

The following questions were asked:

1. Origins and motivation

a. Why did you undertake your S2ST project and lead it to an operational conclusion?

b. What are the immediate goals and achievements?

c. What longer term goals do you have?

2. Technology

a. Which MT system do you use?

b. Which ASR tech do you use?

c. Which TTS tech do you use?
d. Which emerging technologies do you see as becoming relevant?

3. Use case and market

a. What is your primary use case? Are any other use cases emerging?
b. What is your target market? Which cohort of users?
c. What is your business model for this product/service?
d. How do you price the product/service?

4. Language pairs

a. Which language pairs are most used/required?
b. Which new ones do you plan to develop and why?

5. SWOT analysis

a. Strengths (best use cases for S2ST: your organization's strong points)
b. Weaknesses (worst use cases for S2ST: your organization's weak points)
c. Opportunities (confluence of technologies, needs, and lifestyles)
d. Threats (e.g., a disruptive technology shift)

Below, we'll present the interviews alphabetically by organization. For smoothness, we summarize rather than quote the responses. To avoid redundancy, we include only selected responses, but these have been reviewed and approved by the respondents.

NOTE: The opinions and claims of interviewees are their own, and in no way reflect the policies of TAUS. In this interview context, the same caveat applies to the personal statements of interviewees Mark Seligman and Alex Waibel, also co-authors of this report.

Alibaba (Eric Liu)

Origins and motivation: Alibaba has no full-time SLT team: I'm part of the Smart Engineering team, of which the main role is to improve service for the ecommerce business and for our cloud users. We do have a business collaboration software app, and this will become the basis of work on SLT.

Our current focus, however, is not on operational translation engineering, but on infrastructure and technology development.

In these efforts, we're facing some current bottlenecks. ASR is improving, but MT is not yet satisfactory, and this is the major barrier. The total experience isn't yet "magical." We judge, for instance, that Skype Translator isn't yet ready for consumers. So we expect to develop more aggressively when the time is right.

Technology: We have own systems for all of the major components – ASR, MT, etc. Our main focus remains upon the quality of the user experience.

Use case and market: Our ideal use case would be business communications, e.g. for Chinese businesses selling abroad, as opposed to general consumer use. These businesspeople are currently using our text translation, so an eventual shift to speech can be expected, with the development of new habits. At the present, however, there is no strong demand.

Language pairs: To date, we concentrate on Chinese and English as our main language is of interest.

Eric Liu, General Manager, Alibaba Language Services

Eric Liu is the General Manager of Alibaba Language Services, the solutions and services division within Internet giant Alibaba Group responsible for integrating language crowdsourcing, big data and monetization into the globalization efforts of a wide range of Alibaba business units, including Tmall, Alibaba.com, AE, Alipay, Aliyun, AliTrip, Dingtalk, Youku and UC.
Eric joined Alibaba Group in 2015 when Alibaba acquired a crowdsourcing technology company he founded. Eric Liu is a graduate of Vassar College in New York.

Alibaba Group, China, 1999
Number of employees: 46,228
http://www.alibabagroup.com
S2ST product: Not applicable

Chinese Academy of Sciences (Chengqing Zong)

Origins and motivation: We've been developing our technology as researchers for many years now. Originally, the topic was text translation, but I wanted to push in the direction of speech translation.
One major project related to helping tourists coming to the Beijing Olympics in 2008. In the end it was not used, however: there were many volunteers who could translate, and system performance wasn't really good enough. However, many of my colleagues are still carrying on speech translation research, for example for Chinese travelers who go abroad.

Technology: Most of this research employed statistical machine translation, but this year we have begun to experiment with neural networks. This research is still in the early stage, though. Presently the technology is too slow for real-time use, so phrase-based SMT is still the technology of choice. How should the speed problem be addressed? We don't know! We don't have a good strategy yet. However this is a widespread problem, not ours alone. In any case, we always use our own technology – never commercial or third-party tech, either for translation or speech elements.

Future directions may include different means of evaluation. Until now, we've been using automatic techniques such as Bleu Scores. In the future, we'll rely more on human judgment. Work on neural networks will continue, despite problems with speed and data sparseness. We're also studying how to combine neural networks with more traditional techniques. Another pressing concern is the treatment of out-of-vocabulary – that is, unknown – terms.
We have not been researching specialized devices, since we think it's more practical to use devices like mobile phones that everybody has.

Use case and market: I think that, for speech translation, the best use case is still for travelers. Other use cases, such as transportation at conferences, pose various problems, for instance with specialized terms. We haven't considered military or humanitarian uses yet, but they seem difficult and impractical because the situations are so complex.

Language pairs: The languages we work on include English, Chinese, Japanese, and minority languages in China such as Tibetan and Mongolian. There are more than 10 language pairs, all to and from Chinese. The minority languages are useful, since the speakers often cannot speak good Chinese.

Prof. Chengqing Zong, Research Fellow, National Pattern Recognition Laboratory

Chengqing Zong received his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), in March 1998. He is a Professor at the National Laboratory of Pattern Recognition, CAS's Institute of Automation. His research interests include machine translation and natural language processing. He is a member of International Committee on Computational Linguistics. He is associate editor of ACM TALLIP and editorial board member of IEEE Intelligent Systems, Machine Translation, and Journal of Computer Science and Technology. Also, he served ACL-IJCNLP 2015 as a PC co-chair, COLING 2010 as an organization committee co-chair, and many other top-tier international conferences, such as AAAI, IJCAI, WWW and CIKM etc., as Senior PC member, PC member or other roles.

Institute of Automation, Chinese Academy of Sciences, China, 1956
Number of employees: 566
http://english.ia.cas.cn/
S2ST product: Not applicable

CMU/Karlsruhe Institute of Technology (Alex Waibel)

Origins and motivation: Our research has always been motivated and driven by a desire to contribute to human cross-cultural understanding and communication. In a globalizing world, language is one of the most significant barriers generating separation and misunderstandings. Human interpretation, while still much better, is not always available or too costly; thus there's an enormous

need for automatic solutions. However, in our view, the problem is not translation of text, but cross-cultural understanding. As such, the problem is much broader, and we realize that it involves speech, gesture, text, images, facial expression, cultural and social understanding, emotion, emphasis, and many other cues and sources of information. Along this journey, our lab has been particularly active in researching the problem of speech translation, as it harbors many of these cues and human communication elements. While working on solving many of the underlying technical problems, we have also been active transitioning the results to society by way of start-up companies, services deployed for universities and governments, and engagement in humanitarian deployments. With better technology, we've been able to support better communication tools in the wild, and data from those deployments have informed and motivated additional research.

Technology: Our first speech translators (which were, in fact, the very first in the US and Europe) began with a mix of rule-based, statistical, and neural processing and thus have always built on the current state of the art. As a scientific vision, however, we have always believed that any usable and scalable solution must build on machine learning and adaptation, to capture the enormous ambiguities, complexities, and contextual interactions of human language. During each phase of our research and development, we also had to struggle with the available resources of the day to build the best possible solutions. Initial rule-based translation could be replaced by statistical and neural methods as computing and data resources grew. In speech recognition, too, we began with speech translators that already ran with deep neural networks in the late 1980s.

In 1987, we proposed the Time-delay Neural Network (TDNN), the first convolutional neural network, and applied it to speech recognition. The TDNN delivered excellent performance in comparison to HMMs [Waibel, 1987; Waibel et al., 1987]; but, due to lack of data and computing power, the superiority of neural network approaches over statistical methods (which were more popular at the time) could not yet be demonstrated conclusively. With several orders of magnitude (1000 to 10,000 times) more computing and data, however, the same models now generate up to 30% better results relative to the best statistical systems. This development and similar advances in machine translation have now finally led to a widespread shift to neural network methods in the community as a whole. A combination of better algorithms with faster and more powerful computing was also a key enabler for the types of systems and deployments we could realize. At first, our systems had to be limited in vocabulary and complexity to make them work, but we're now able to deploy entire interpreting systems with unlimited vocabularies, operating in real-time and low latency on lectures, and even running (somewhat reduced) versions on smartphones.

The community also had to overcome its biases and fears. Machine learning, as a solution to speech and translation problems, was initially fiercely attacked and criticized by our colleagues, because one could not "understand" what the system does. This issue was epitomized by neural networks, where we cannot even describe or model what the internal neural nodes learn. We were often described as neural "nuts" and the research criticized as "unscientific." This criticism never bothered us, since our very own brains don't understand how we do things; and though we can rarely describe or define precisely why and how we call something a "chair" or a "table," we can recognize them exceedingly well. Now we know that machine learning does indeed yield powerful solutions, leading to the recent renaissance in AI and "Deep Learning." Our work has extended to human interface factors as well. As just one example, we first proposed the presentation of translated "subtitles" on wearable devices or heads-up display goggles during conversations [Yang et al., 1999]. The technology was conceived to provide translation in mobile situations. A variant of this idea was demonstrated as an application in Google Glass[1].

Use case and market: Early on and throughout our research, tourism and medical exchanges have been fruitful use cases. Initially, we could

1 See e.g. https://www.cnet.com/news/real-time-real-world-captioning-comes-to-google-glass/.

limit apps to specific "domains"; in the early '90s, we envisioned early pocket translators for tourists and healthcare workers. A second application would be for humanitarian missions, to permit non-experts without knowledge of a language to dialogue effectively in the field. And with these users in mind, we also aimed for embedded systems to be used off the grid – good and fast, but not necessarily networked.

Beyond research systems and milestones, our start-up company Jibbigo successfully built and deployed the first-ever speech translation app that ran in real time over a 40,000 word vocabulary, without the need for network access. It was used in a variety of humanitarian missions, but was also sold to travelers via the iPhone and Android app stores. Jibbigo was acquired by Facebook in 2013, where the team continues to build speech and translation products. The speech translation use case in mobile environments continues to provide opportunities, and a number of players now produce products for consumer mass markets such as travel and hospitality. These are mostly network-based, using recognition and translation services provided by cloud-based tech companies.

Another area important to us is that of broadcast news. There the idea is to work with big media companies to make content available rapidly in other languages, for instance giving multilingual access to information on YouTube, TV programs, etc. In 2012, I had the opportunity to lead a large Integrated Project called EU-Bridge, funded by the European Commission. The program involved nine top research teams and developed services to develop cross-lingual communication services for a multilingual Europe. One specific use case was the automatic and immediate interpretation of broadcast news. The aim was to produce automatic transcription and interpretation services (via subtitles) for the *BBC*, *Euronews*, and *Skynews*. Deployment activities are continuing since the completion of the project in 2015.

We've pioneered another use case since 2005: interpretation of lectures and seminars and political speeches and debates, in which there's vast potential for minimizing the language barrier. There's also tremendous demand, since human resources are often unavailable or too costly. Given the formidable language barriers in this setting, the question isn't whether technology is better than humans, but whether it's better than nothing.

Technically, for the teacher-to-student direction, we can deliver output as synthetic speech over head-phones and as subtitles on a device. The latter tends to be more practical, less obtrusive, and cheaper during a lecture, so our deployed systems deliver text output via a regular Web browser. The browser can be used on a mobile phone, tablet, or PC, so any student can easily obtain access to the result. Under project EU-Bridge (mentioned above) we were able to expand our work to create an SLT- for-lectures service that is now deployed in many lecture halls at university in Karlsruhe and other cities in Germany, interpreting German lectures for foreign students. The automatic subtitling also provides a solution for the hearing impaired, who can follow along in real time. We are also carrying out extensive field tests using computer speech translators and language tech components at the European Parliament, where we explore interfaces and environments that optimize a joint and synergetic workflow between human interpreters and technology to decrease stress and increase productivity.

Looking toward the future, one major problem stifling advances at present is that large platforms distribute SLT for free as a feature around other services. This makes it difficult for technology companies to thrive and to provide optimized solutions in any space. But significant opportunities continue to exist in vertical markets, including, medicine, education, social and humanitarian services, and government (in interpreting services, disaster relief, law-enforcement, the military, and other areas).

Language pairs: Jibbigo covered 15 to 20 European and Asian languages and pairings between them; and at our research labs we have worked on about 20 languages. But the cost of developing new languages is still too high to address all 7,000 languages of the world. What about Khmer, Arabic dialects, or

provincial languages in Africa or South East Asia? Many of these languages are not well researched, and for many no data can be found on the Internet. Many are only spoken and lack an orthography, and many vary regionally by accent and dialect. For the development of a system, however, extensive speech and translation databases have to be collected; and vocabularies, language models, and acoustic models must be built, trained, and adapted to each language, domain, and task. To get closer to the dream of a world without language barriers, the cost of system creation for a language and use case must be reduced dramatically. Our team has carried out many research projects in search of solutions to this problem. As we progress, many of the current component technologies will gradually become language-independent or language-adaptive, or will collect their own data autonomously. Interactive technologies and crowdsourcing will also make systems adaptable by non-experts during field use.

Another concern is cross-language pairings between languages. When translation technologies switched from interlingua-based approaches to statistical translation – which till now has entailed directly pair-wise language learning – to scale up and gain domain-independence, we lost the nice property of connecting arbitrary languages through an unambiguous intermediate concept representation, or interlingua. The idea, however, is returning through newer, faster, and larger neural network models than were possible in the later '80s and '90s. If it becomes possible to train arbitrary semantic representations across several languages, the addition of new languages should become easier. This is a theme of our current ongoing research at the University labs.

SWOT analysis:

Strengths: Our group has been strong in two areas in particular: first, we have pioneered off-line SLT for smart phones; second, we have introduced the first lecture translator in 2005 and are advancing the state of the art in simultaneous interpretation *services*. We operate the only current simultaneous interpretation service at Karlsruhe Institute of Technology, and are growing a user base at partner institutions. To assist foreign students, the systems translate simultaneously during lectures at very low latencies (that is, little time is needed for the translation to emerge), and also after class in review mode. We're also field testing interpretation and language component technologies to assist human interpreters at the European Union Parliament.

Weaknesses: We still lack real understanding of how semantics works. We're still operating too much on the surface words rather than on the intended meaning. We need better models

Prof. Alexander Waibel, Director of the Interactive Systems Laboratories

Dr. Alexander Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the Karlsruhe Institute of Technology, Germany. He is the director of the International Center for Advanced Communication Technologies (interACT). The Center works in a network with eight of the world's top research institutions. Dr. Waibel's team developed and demonstrated the first speech translation systems in Europe&USA (1990/1991 (ICASSP'91)), the world's first simultaneous lecture translation system (2005), and Jibbigo, the world's first commercial speech translator on a phone (2009). Dr. Waibel founded and served as chairmen of C-STAR in 1991. Since then he directed and coordinated many research programs in speech, translation, multimodal interfaces and machine learning in the US, Europe and Asia. During his career, Dr. Waibel founded and built 10 successful companies. Since 2007, Dr. Waibel and his team also deployed speech translation technologies in humanitarian and disaster relief missions. Since 2012, his team also deployed the first simultaneous interpretation service for lectures at Universities and interpretation tools at the European Parliament.

Carnegie Mellon University, USA, 1900
Number of employees: 1423
http://www.cs.cmu.edu/
S2ST product: Not applicable

of conceptual representation. Porting language technologies to new languages, domains, and speaking styles is still too costly. We need better "portability" and self-maintenance of language systems. We need better models of the conversational context to deduce the underlying meaning – models including the speakers' gender, emotion, their dramatic performance, social setting (chatting or lecturing), social relationships (status, gender, rank,..), etc. Context could also help ASR to resolve e.g. names; help MT to automatically resolve ambiguity; and so on.

Opportunities: It would be worthwhile to integrate MT with multimodal graphics and sensing technology to enhance human communication.

Threats: People consider SLT to be a solved problem! This misconception means lack of funding for advanced research. Or, at the opposite extreme, the problem is still derided as impossible. The truth, of course, is in between: practical, cost effective solutions do exist today, yet a great deal of work is still needed for new and additional use cases. Large companies are helping people to connect across languages, but they're few in number. Thus, to an extent, they pose a threat: they make it hard for small, agile innovators to stay afloat and make progress, and they restrict research by not sharing results and data. And yet it's increasingly recognized that natural language understanding and SLT are crucial to business success.

EML (Siegfried "Jimmy" Kunzmann)

Origins and motivation: Over the past thirty years, I've developed several technologies in the speech area that we now license. Our ASR, in particular, is a core technology for SLT. At IBM, I worked on automatic speech recognition, synthesis, natural language understanding, and translation technologies used in the IBM MASTOR project to facilitate communication with doctors. I also participated in the TC-STAR consortium to develop early SLT systems.

At EML, we're currently cooperating with Lexifone (IS) on a project for smartphones focused upon real-time transcription and translation of financial news. The input data sources and semantics are quite complex. We also work on a wide range of human-computer interaction applications. Overall, our focus for SLT is upon opening applications and services to additional human-human and human-machine domains exhibiting recognition and semantic problems.

Technology: We provide our own ASR technologies to our partners for installation in enterprises and the creation of customer services. We offer a broad range of APIs for integrating our ASR into products. All of these are designed for on-premises installation – for companies that care about private and company data security and want to work in their own cloud environment. Because many companies are now giving smart phones to their staff, we expect ASR usage to spread widely and think that SLT technologies will be used on-premises for e.g. mail and Web-services. SLT apps for medical and business applications also require data privacy, which again implies on-premises deployments. We work with the German supplier Linguatec on integrated MT technology solutions. As with MT technologies, we partner for text-to-speech technologies with other, usually European, companies. This is our usual procedure: working in partnerships to bring SLT solutions to the market.

Use case and market: We're interested in the enterprise market for scalability, and this involves privacy and company data security issues. Within this area, there is the enterprise messaging or texting market. There's a demand to have speech recognition and machine translation running on the local intranet rather than on the larger Internet outside the enterprise, say with vocabularies of more than one million words. These facilities can then be used for dictating letters, for voice search, and for translation as an all-in-one service.

One attractive area is transcription and translation of telephone conversations. We've been working with Lexifone in this area. There should be many more speech-to-text applications for enterprise communications, with different components arranged in various ways for various services. However, phone bandwidth is limited, so accurate performance must usually be achieved by customizing the domain vocabulary with appropriate tools.

Another speech-intensive area is that of speech analytics, in which e.g. the call performance of customer and agent conversations is analyzed and evaluated. Until now, we've done such work offline – that is, on recorded content – but we anticipate extensions to real-time analysis soon. Since such analysis usually calls for full transcription, the transcripts can also be fed into MT when necessary.

The analysis of phone conversations via ASR is closely related to comparable analysis of media, such as broadcasts or podcasts. And the same speech recognition technologies used for analytics can be used to generate close captioning and/or subtitles. The necessary ASR can be done off-line or in real time. Presently we're working with live subtitling and transcription for assistive service (e.g. for deaf people in conferences etc.). Similar work relates to live transcriptions for online courses. Business models depend on the partners and their markets. In call centers, for instance, the payment is one-time fee-based, plus a maintenance charge. For text messaging, charges are time-based on volume of audio – that is, on minutes used, or per user, or per message.

Language pairs: In line with our need to compete with large companies, we currently cover essentially all European and North American languages, plus Arabic and Chinese. As we start to address the automobile market, we'll be offering the 12 first-round languages. For Asia, we'll plan to add Japanese and Korean as well.

SWOT analysis:

Strengths: Our strongest business is recognition and transcription of voicemails for the telco market. There are many customers and a huge quantity of data throughput.

Weaknesses: Voice search is the weakest market – and this is ironic, since this area has currently the biggest potential! The profit is weak because large platforms presently give voice search away for free. We assume that the business sector may invest here, but, from the perspective of revenue, there is not yet much revenue here. However, a large amount of data is available.

Opportunities: We envision a great potential market for communication with, and control of, business smart phones.

Threats: A closed breakthrough technology solution would be dangerous for us as a small company. However, our license income based on state-of-the-art ASR technologies helps to protect our business, as does our partner network. A more general R&D threat is the lack of funding in the European Community for SLT technologies.

Dr. Siegfried "Jimmy" Kunzmann, R&D Manager

Since 2006 Dr. Siegfried ("Jimmy") Kunzmann has been R&D manager of the EML European Media Laboratory GmbH (Heidelberg, Germany). The speech transcription platform for server and local decoding uses machine learning techniques to support voicemail-to-text, mobile messaging & search, speech & media analytics, broadcast news subtitling and car, house & media voice control applications. From 1991-2006 he managed IBM's European Voice Technology organization. It focussed on multi-lingual speech processing, language technology tools, and techniques for language specific needs. Jimmy Kunzmann holds a diploma degree in Computer Science and a PhD in speech processing from the University of Erlangen-Nuremberg, published more than 40 papers, is author of one textbook and filed more than 10 patents.

EML European Media Laboratory GmbH, Germany, 1997
Number of employees: 15
http://www.eml.org/english/index.php
S2ST product: EML Transcription Platform

IBM/Microsoft Applied Sciences Group (Yuqing Gao)

Origins and motivation: I worked on IBM's MASTOR project in the late 2000s, due to my background in ASR, natural language understanding, and TTS. The mission was straightforward: to break the language barrier. We had a small team of six on an internal IBM project

funded by DARPA; by 2007, the team grew to twenty. The initial focus was on SLT between English and Iraqi Arabic. A dedicated handheld device was used as the platform, giving challenges with respect to size, memory, and noise. Back-translation was provided as feedback to users on translation quality, using an interlingua for the semantic representation. We were able to leverage this representation to compensate for limitations in training data. The main use case was humanitarian, e.g. for medical conversations, interrogating locals, etc. In 2007 the device was donated to the US government and 1,000 dedicated hardware devices were used in Iraq. Ten thousand software licenses were donated by IBM.

In 2008, IBM began to look at commercial applications for the text translation component, with up to 13 language pairs. Hosted as an internal IBM service since August 2008, n.Fluent offers a secure real-time translation tool that translates text in Web pages, electronic documents, and Sametime instant message chats. A BlackBerry mobile translation app uses the same software.

Technology: n.Fluent has been hosted as an internal IBM service since August 2008. The software offers a secure real-time tool for translating text in Web pages, electronic documents, Sametime instant message chats, etc. A BlackBerry mobile translation application is available. Languages handled are English to and from Arabic, simplified and traditional Chinese, French, German, Italian, Japanese, Korean, Portuguese, Russian, and Spanish.

Use case and market: Until 2008, work on MASTOR's SLT concentrated on healthcare. Subsequently, work was extended to military and travel.

SWOT analysis:
Strengths: During the development of the MASTOR project, it was an advantage that we were targeting a handheld device. It was also helpful that we used semantically informed and interlingua-based techniques as compensation for the lack of extensive data. And we performed a lot of domain study, for instance in healthcare.

Weaknesses: We needed to cope with problems related to computing power and sparse data, as was normal for the times. Sub-par audio systems were also an issue.

Opportunities: We saw many openings in the travel domain, e.g. for elderly people finding themselves in foreign countries. Chat services also seemed promising. And we saw the chance to combine semantics with MT and natural language understanding in interesting new ways.

Threats: We felt that the hope to develop general-purpose speech translation, usable in any situation, was misguided, and in fact a liability to the field because it prompted false expectations. It's best to aim instead at SLT for specific use cases.

Dr. Yuqing Gaom, Microsoft Partner, Group Engineering Manager (GEM)

Dr Gao is an accomplished scientist, innovator and R&D leader in cutting-edge research and product development. She has a proven track record of success (23 years of successful career in Microsoft, IBM and Apple), and broad range of skills from research to product development. Her expertise includes enterprise middleware, cloud computing, workload optimization, business analytics, big data, speech recognition, NLP and ML. Her work was featured by MIT Technology Review, Time, CNN, ABC, BBC, etc. She is an IEEE Fellow for her distinguished contribution to speech recognition, speech-to-speech translation and natural language understanding. She published over 120 papers, holds 35 patents. She was an IBM Distinguished Engineer, and now is a Microsoft Partner, and GEM for BING Knowledge Graph.

Microsoft, USA, 1975
Number of employees: 120,000
http://www.microsoft.com
S2ST product: IBM MASTOR

Lexifone (Ike Sagie)

Origins and motivation: Lexifone provides telephone-based speech translation. Our current orientation is mostly toward business rather than consumer use, e.g. for travel. Presently, the main use cases involve calls to businesses and call centers. When foreign language calls arrive at businesses (e.g. hotels), Lexifone's service can be brought in as an intermediary. Both sides are informed that an automatic interpreter is operating.

We also offer service to call centers. Both human and automatic interpreters can be brought in: automatic interpreting can function alone, with considerable cost savings (though humans can be brought in for clarification or correction); or it can function alongside humans to assist, e.g. by showing a text transcription of the call. While our emphasis is not presently on consumer interactions, we claim that we could handle these with higher quality than either Google or Skype.

Technology: You can't simply use existing engines for machine translation and speech recognition. Optimization layers and other modifications are also required. First, people speak continuously, so there must be an acoustic division solution that cuts the flow into sentences or segments and sends the output to an audio optimization layer. Linguistic optimization is then needed in the next stage to ensure the translation accuracy, for instance to make sure interrogative sentences are annotated with question marks.

Language pairs: In our business-oriented space, the top languages to date are English, Spanish, Portuguese, Mandarin Chinese, and Cantonese Chinese, in all combinations. We presently offer over 17 languages in all combinations. We have seven new languages in beta, including Japanese.

SWOT analysis:

Strengths: We've taken a multidisciplinary approach to improve the overall user experience – acoustic and linguistic optimization, combination of these technologies with video, etc. We feel this approach yields a unified technological solution.

Weaknesses: We're a small company, and need investment to compete with the big boys. Concerning weaknesses of SLT overall: it's not yet clear that the market will accept MT replacing human interpreters. For example, how will users of the call center react to an automatic solution when agents cannot speak their language? We need to explore the possibility of mixing human and machine interpreters. In the business fear, concern with security also poses an obstacle to acceptance.

Opportunities: We hope to become a leader in SLT for business use. This is a multibillion-dollar market, but we're still only at the gadget stage.

Threats: If big players decide to address the business side of the market, it could spell defeat for our focus. We hope nevertheless to build our technology into the phone and device maker market.

Dr. Ike Sagie, Founder, CEO

Dr. Ike Sagie has over 30 years of computer science and computational linguistics experience. He previously founded and served as the Chief Technology Officer of Attunity (Nasdaq: ATTU; Market Cap $76M), a provider of information availability software solutions. Prior to founding Attunity, Dr. Sagie served as the head of the Software Engineering and Programming Languages department at IBM. He earned his PhD in Computer Science from Technion-Machon Technologi Le' Israel.

Lexifone, Israel, 2010
Number of employees: 6
http://www.lexifone.com/
S2ST product: Lexifone in-call interpreter

Logbar (Takuro Yoshida)

Origins and motivation: I personally experienced the strain of communication difficulties when learning English while living overseas, and this prompted my own interest in translation technology. Later, in 2013, I founded

Logbar in Tokyo and London as a "social bar" in which everyone had an iPad for communication. In this social context, the need for translation was clear. We got some media coverage, but gave up after six months.

Our next project was development of the Ring: you wear it on your finger and use it to control various devices through hand motions. This effort cultivated our expertise in engineering small computational devices, as needed for our speech translation project.

Many translation applications depend on a Wi-Fi connection, a drawback for travelers. So in 2015 we began to develop a portable SLT system requiring no Internet. It's worn as a necklace, and will be used mostly hands-free to provide the simplest possible cross-lingual communication. We expect to launch in Japan, China, Taiwan, and the US in late 2016 or early 2017.

The device's name, "ili," graphically represents two people talking across a barrier.

Technology: Our technology for ASR, MT, and TTS is licensed from the NICT research organization; we work closely with them. For our product, a translating necklace usable offline, we had to fit everything onto the device, so we had to shrink a server-based system down to a small footprint in terms of CPU etc. Our operating system is proprietary, and created in-house for speed and agility. We don't want to say too much about our interface at this point, but it's also programmed in-house, and partly button-driven.

Use case and market: Our translating necklace product was initially intended for social use, for instance in bars. Going forward, we'll focus on the travel industry and more generally on consumer communications, for instance in Japan and China. We're not yet targeting military, healthcare, or customer services.

Our business model will initially be to sell the necklace product as hardware. A little further along, we're thinking about B2B systems as a set of services, and these will be sold as licenses. Pricing models are still to be decided for all of these use cases.

Language pairs: Since we presently depend on NICT technology, we're constrained by the languages that they offer. Initially, however, we'll certainly include Japanese, Chinese, and American English. Then we'll move on to other languages as our experience in the growing market dictate.

SWOT analysis:
Strengths: We're good at making small-scale devices with small energy requirements. So we're interested in saving energy for wearable devices, for example with small batteries. Another plus is that we're both a hardware and software company. Since we make stand-alone devices, our focus is on offline use, connecting to the cloud only when absolutely necessary.

Where machine learning is concerned, as needed for image and speech recognition, our emphasis is on alignment with consumer needs, so everything has to be cheap and small-scale. For consumers the key is functionality, not technology in itself. We're using NICT technology for speech and translation, but hope to add value to it through our refinements for practical wearable use.

Weaknesses: We're aware of the need to improve our marketing, especially since we're still small. In particular, we shouldn't market the necklace device as a translation system! Instead, we should emphasize its functionality so as to address real desires. Again: a desire, not a technology. Pushing magic is a weakness.

Opportunities: The whole world needs this sort of functionality; so, going forward, we can focus on new language pairs and contexts.

Threats: There's a harmful and unreasonable tendency to feel that if translation isn't 100% correct, people won't use it. But in fact we're already using less-than-perfect translation. Imperfect accuracy for a self-driving car should worry people, but a translator is different.

Takuro Yoshida, Founder & CEO

Logbar, Japan, 2013
http://logbar.jp/
S2ST product: ili

Microsoft/Skype (Chris Wendt)

Origins and motivation: The time was right for speech translation: a number of factors converging in a perfect storm came together over the past five years or so. ASR underwent dramatic improvements by applying deep neural networks; MT on conversational content had become reasonably usable; and Skype provided an audience already engaged in global communication.

Technology: ASR, MT, and TTS by themselves are not enough to make a translated conversation work. Clean input to translation is necessary; so elements of spontaneous language – hesitations, repetitions, corrections, etc. – must be cleaned between ASR and MT. For this purpose, Microsoft has built a facility called TrueText to turn what you said into what you wanted to say. Because it's trained on real-world data, it works best on the most common mistakes.

Another necessity is a smooth user experience. We know that MT isn't perfect. How can we guide speakers and protect them from errors and misuse of the system? Our first thought in this direction was to model the interpreter as a *persona* who could interact with each party and negotiate the conversation as a "manager." But we decided against it for several reasons.

First, a simulated manager doesn't always simplify. On the contrary, it can make the content even more ambiguous by drawing attention to possible problems and adding additional interaction, so that the experience becomes too complex.

Next, when we watch people using the system, they tend to forget the interface and speak normally. And so, in the spirit of Skype, we want people to communicate directly with each other, keeping the technology behind the scenes. The translation should appear as a translated version of the speaker, with no intermediary. In fact, we note that people who use human interpreters professionally tend to ignore the presence of the interpreter. They pace themselves and allow the translation to work at its own pace. Parties to the conversation focus on the speaker, not the interpreter. By contrast, novices tend to talk to the interpreter as a third person. So we decided against using the interpreter idea. In Skype, you see a transcript of your own speech and the translation of what the other person said. You can also *hear* what the other person said via speech synthesis in your language. But in fact, in our usability tests, only 50% of the users wanted to hear translated audio – the others preferred and found it faster to hear only the original audio and follow the translation by reading on screen.

Another interface issue relates to the relation between the original voice and the spoken translation. The most straightforward procedure is to wait until the speaker finishes, and then produce the spoken translation. But to save time and provide a smoother experience, we are now incorporating a technique called *ducking*. It's borrowed from broadcasting: you hear the speaker begin in the foreign language, but the spoken translation begins before he or she finishes, and as it proceeds the original voice continues at a lower volume in the background[2]. The ducking is intended to encourage speakers to talk continuously, but most people I've talked to still wait until the translation has been completed. Unlike a text translation, spoken translation cannot be undone; so once the translation has been finalized, the speaker will hear it. This factor poses difficulties for simultaneous translation, since segments later in the utterance might otherwise affect the translation of earlier segments.

Use case and market: Microsoft offers the speech translation service via a Web service API, free for up to two hours a month, and at a base price of $12 per hour for higher volumes, heavily discounted for larger contingents.

Language pairs: We presently handle nine languages for SLT, in all directions, any to any.

Arabic can presently be paired with Italian, English, French, German, Spanish, Portuguese, Mandarin Chinese, and Russian.

In addition, we can handle multidirectional text translation for 50 languages, and speech

2 In this way, the Microsoft-Skype SLT system now includes elements of simultaneous translation. – Editor

translation from the 9 languages to 50 others.

SWOT analysis:
Strengths: We feel especially strong in the area of general consumer conversation, e.g. for family communication, as when for example grandmothers can speak across language barriers to their overseas grandkids.

Weaknesses: Names pose particular problems for us. It just isn't possible to list in advance all names in all of the languages we handle, either for speech recognition or for translation.

Opportunities: The quality of MT is continuously increasing, and with it the range of scenarios in which the technology can have a positive impact. We see that the tech finds its applications almost autonomously: the applications match the achievable quality level. At the Ignite conference, Microsoft showed lecture translations, one speaker to a large audience, in Skype for Business.

Threats: Speech as a medium of communication may be losing out to text for some use cases. Young people tend to IM or to use Snapchat more than to phone each other, for instance. So the demand for speech translation could suffer in these scenarios. On the other hand, we expect increasing interest from enterprises, e.g. those active in consumer services.

Chris Wendt, Principal Group Program Manager

Chris Wendt graduated as Diplom-Informatiker from the University of Hamburg, Germany, and subsequently spent a decade on software internationalization for a multitude of Microsoft products, including Windows, Internet Explorer, MSN and Windows Live - bringing these products to market with equal functionality worldwide. Since 2005 he is leading the program management and planning for Microsoft's Machine Translation development, responsible for Bing Translator and Microsoft Translator services. He is based at Microsoft headquarters in Redmond, Washington.

Microsoft, USA, 1975
Number of employees: 114,000
https://www.microsoft.com
S2ST product: Skype Translator

NICT (Eiichiro Sumita)

Origins and motivation: We began as a research-only organization, but now sell speech and translation technologies to private companies in Japan. The history of NICT (the National Institute of Communication and Technology) goes back to that of ATR (Advanced Telecommunications Research) International, which itself has a rather complicated history.

In 1986, the Japanese government began basic research into machine translation. Because the Japanese language is unique, it was felt that Japanese people need MT systems to enhance their lives. At the time, however, this effort appeared as Mission Impossible: little practical technology was then available. Consequently, this R&D area appeared appropriate for government sponsorship. This research became the source of the ATR project, which added speech translation to the goal set. There was no commercial goal at that time. ATR 1 was followed by ATR in 1992. Then, in 2008, a group of ATR researchers moved to NICT. A more accurate system, presented in 2010, drew the attention of many companies, which then created their own speech translation systems. The DoCoMo system was particularly prominent.
However, NICT is still completely sponsored by the Japanese government. We'll continue to carry out research, organized in five-year plans. We're now working toward 2020, aiming at a system to be used in the Olympics. Our job is to provide the basic technology, so that private companies can produce their own systems and products for the 2020 events. We'll provide not only APIs for the three main components – ASR, MT, and TTS – but optical character recognition, noise reduction, and more.

Use case and market: We're planning to put this translation system into use for hospitals, accident response, shopping, and many other community functions. We want to handle basic conversations for daily life. The business model depends upon the company and product which incorporate the foundational technology that we provide. It could depend upon advertising

or licensing, for instance. One current example is DoCoMo's Hanashita Honyaku product. Another is the Narita airport translator. A third is the effort by startup Logbar to develop its special wearable hardware for standalone translation. Perhaps Panasonic will develop an SLT device. With respect to special-purpose versus general-purpose devices such as smart phones, I think special devices are a very good idea. We should indeed develop devices suitable for SLT. For example, there should be hands-free devices for people in hospitals who can't use their hands.

Language pairs: We're presently concentrating on 10 languages, including Japanese, Chinese, Korean, Indonesian, Thai, French, Spanish, and Myanmar. We'd like to do more, but our budget is limited. Free downloads are available on app stores for a product named VoiceTra which handles all of our language pairs. As an unusual feature, we don't use pivot languages to handle all of the possible language paths; instead about 900 direct paths have been developed.

SWOT analysis:
Strengths: Our strongest use case is in communication for tourism. Our quality is higher than that of the Microsoft or Google program within this domain, especially in the hotel service area. Overall, though, we expect a market of 40 billion yen by 2020.

Weaknesses: Compared with Google, we're not yet equipped for general-purpose systems. We hope to improve, but doubt we'll be competitive in this area by 2020. There are some technological lags, too: we haven't yet tried to combine statistical and neural network approaches, for instance.

Opportunities: We plan to develop SLT systems for translation of broadcasts, e.g. for news. So again, development of general-purpose systems will be important.

Threats: The only real threat is lack of money! The R&D budget for SLT in Japan is not big. The country has a lot of debt.

Speechlogger (Ronan Rabinovici)

Origins and motivation: My interest in translation started with difficulties in communicating with my grandparents, either by telephone or face-to-face. I work in high-tech, so it was natural to feel, "Let's try to do something!" We began in 2014. Since then, we've been getting feedback from users all over the world and doing things for them. We have a free application now that does many things – maybe too many. I'm a mathematician and physicist, so we made it perhaps too complex: transcription, speech recognition, punctuation, and translation into several languages.

As a result, Speechlogger doesn't have really good reviews – we think because it combines four rather difficult functions. All of these technologies are in their early stage, so when you combine them, you get something imperfect. And the market doesn't really need a complex system now. It turns out that most of our users just want transcription. So we have spun off a new dictation effort called Speechnotes, and it has become the most successful Web-based ASR application on the Web. Unlike Speechlogger, it's very simple. It does dictation only, but it does it very well.

Meanwhile, Speechlogger is still operating and we will improve it, but it's not our host

Dr. Eng. Eiichiro Sumita, NICT Fellow

NICT Fellow, Dr. Eng. Eiichiro SUMITA received the M.S. degree in computer science from the University of Electro-Communications in 1982 and the Ph.D. degree in engineering from Kyoto University in 1999. He is the Associate Director General, ASTREC (Advanced Speech Translation Research and Development Promotion Center), NICT. Before joining NICT, he worked for ATR and IBM. His research interests include machine translation and e-Learning.

National Institute of Information and Communication Technology (NICT), Japan, 1896
Number of employees: 1,000
http://www.nict.go.jp/en/
S2ST product: VoiceTra

application. Speechnotes is, and it's rated very highly: 4.75 on the Chrome Store.

Technology: We use Google technology for all of the speech translation components – ASR, MT, and TTS. In using Google's ASR, we add patent-pending features to make it more useful and accurate. It's interesting that Google has taken the next step in commercializing their APIs, in particular for speech recognition. Until now, Google's ASR was available only for Android or Chrome, but now you can use it on every machine. You have to pay for it now, but because it is paid there's now a guarantee of service. That's good for a company like us, and for the industry as a whole. In effect, Google is now competing in the speech recognition space with Nuance and IBM.

We have done some internal quality evaluation for speech recognition. Those results aren't ready for publication, but qualitatively speaking, the word error rate is about the same among the major players. Google is certainly far better than any open source ASR that I have tested; and, as you can see on the Speechnotes website, many users of our Google-powered ASR prefer it to Dragon. As for Microsoft, I haven't tried using it for two years; but at that time it was no competition for Google.

Use case and market: All of our applications are presently free (though we ask for donations). Our plan is to teach our users what's possible and then to sell premium services. In charging for ASR, we note that the providers all charge by the minute, but we hope to have some formula that will let us charge monthly. We're also interested in the telephony market, and there they charge per time. So these two use cases are different: for the online market, a monthly fee, and for the communications market, charges per minute.

We do expect a boom in the communications market. For example, at some point every call to a call center will be not only recorded but also transcribed for later analysis. We're working toward that. I think it's only a matter of time. One problem is that the field is kind of crowded.

The technology is difficult in this area, so a very few tech leaders – Google, Apple, IBM, Nuance – dominate the field. It's impossible to compete with them. They make money because they sell their core knowledge as a service. A second problem is that the current level of accuracy is not that great.

So the bottom line is that it's hard to charge for these technologies. That said, for a small company, there are many things left to do or improve. For example, we work on hearing aid devices, cooperating with hardware companies. And there's lots of potential in the field of language learning.

These technologies are awesome for that purpose – to some extent, they can replace a tutor! Not completely of course, but they are better than nothing, especially in developing countries. Google Translate is going in the direction of speaking with people in different countries. The quality isn't great, but it's good enough for many use cases. I don't think it's good enough yet for the professional market, for instance in court. But as the demand increases, the technology will improve.

Language pairs: We use Google technology for translation and speech, and I rarely get a request for a language that I cannot support. It's good that they support so many. For instance, our services support Hebrew, a language with only a few million speakers.

SWOT analysis:
Strengths: Our ASR offering is free! That's not a strength in the product itself, but it is a strength in the market. We have the only free product that works well. Secondly, we're unlimited: we don't put out a limited version to try to upsell; instead, we give you the full product from the start.

Weaknesses: In Speechnotes, the number of features and the design are not so great. Our design should be more modern and more appealing.

Opportunities: For Speechnotes, we should give more features in order to become a paid service. Later, we'll look toward enterprises and the communications market. And as for our plans concerning speech translation ... Well, everything we do is an experiment to test the

market. We like responding to feedback from users. We're having fun. At the moment, I think that the market potential for speech translation isn't big enough for a small company like us. It's okay for large enterprises, but for small companies it's about having fun, not making big money.

Threats: If the big guys decide to give comparable services to consumers, then they will eat us – and no one will hear about Speechlogger anymore! For now, however, they're giving their services to enterprises which can make consumer applications. So, for consumers, we're still great.

Ronen Rabinovici, Founder

Since I was a kid, mathematics was my passion. I took high places in national math competitions. At the age of 14 I was already working for a hi-tech company doing various simple technical jobs. Later I became a teaching assistant in the local math club. This passion to mathematical problem solving is in me to this day. This passion is what led me to studying Math and Physics. After I graduated from the Weizmann Institute of Science in 2008 I worked in research and development for various companies, till I decided to start my own independent path in 2014. Since then, what leads me is producing creative solutions to common every-day problems. Solutions that do good for humanity.

Speechlogger, Israel, 2014
Number of employees: 3
https://speechlogger.appspot.com/en/
S2ST product: Speechlogger

SpeechTrans (John Frei, Yan Auerbach)

Origins and motivation: Both of us have mixed language backgrounds. Yan is from Moldova, near Ukraine (and near Chernobyl). His family came to the US as refugees when he was three. His dad was a veterinarian but had to start all over again here, so he's seen language issues every day of his life. So you could say that our purpose is to help people going through that transition. It's an ongoing mission. The concept is to give our technology away to people who need it – like service agencies and victims of the Japan tsunami – and to have that paid for by people who can afford it. SLT should be available on demand 24/7, and we shouldn't have to wait 50 years for that. At the same time, we don't want to reinvent the wheel. Instead, we want to partner with great companies.

Adopting this broader mission has taken some time. We began in the travel niche as a mobile app. Within six months, we were brought into the Pentagon to bid for a multi-hundred-million-dollar RFP. We didn't get that contract, but we learned a lot. And we finally did sign with Hewlett-Packard, Intel, and Microsoft.

Yan caught the entrepreneurial bug while working as a systems administrator for a venture capital firm while he was still in college.

John is half Swiss, and his early life brought him back and forth between German and English. At one point, his sister-in-law was using Rosetta Stone to learn German to visit Switzerland – but three different languages are spoken there, and she couldn't learn all of them! So he realized that it would be ideal if we could use technology to bridge the language gap. He started working on a proof of concept, cobbling together ASR from Dragon, translation from Google, and a text-to-speech system. When he heard the correct translation spoken, that was the eureka moment for him. He filed for a provisional patent in June 2009 and incorporated in March of 2010. Then in about six months John and Yan built an app for the Apple store. Apple wound up featuring it in several countries.

Technology: Our solution has evolved into a complex combination of technologies. There is a spiderweb of logic behind each API request – and we have about 20. We have translation memory, and if that doesn't work we go to statistical MT, using different engines for different languages and situations. For instance, there's a small company in Saudi Arabia that provides our Arabic; and there are many such partnerships. We collect information from users for continual updates.

We also do post-processing of ASR and MT to enhance accuracy, using correction tools. We think of this revision as providing a basic sort of natural language understanding stage.

We initially built our APIs for Hewlett-Packard, for use in their MyRoom video chat. Later we opened them up so that anybody could use them. For instance, you can build and use your own translation memory.

We've adopted an API-based core business model to enable flexible use of our various technologies for almost any use case. The integration tool we work through now is Zapier, which allows scripting that can coordinate thousands of programs without requiring any programming. Via Zapier, you can use our tools in conjunction with Outlook, Word, Salesforce, and many others.

Emerging form factors are important to us. In China, holding a phone up to someone as part of a translated interchange may be threatening, and a wristband or watch may be better from that viewpoint. We offer both devices and you can control them with your voice. The watch now uses the Android 4.4 operating system. It has its own SIM card, so you can use it separately, without tethering it to your smart phone.

We also offer wireless earbuds, so that you don't need to use the speaker on your device. The earbuds are literally like the Babelfish. You and your conversational partner can each have one. You control them with your voice, for instance to switch languages.

Use case and market: We want to disrupt the entire translation industry across all markets and verticals! We're opportunistic, and there are always new opportunities. For instance, it took us three years to get the Hewlett-Packard contract, but now they keep coming back to us with new openings. Recently, they wanted to compete for a contract at Lincoln Center in New York City. By law, the Center has to provide closed captioning for people who are deaf or hard of hearing. So they called in Sony, but that company's solution would have required replacing the Center's equipment with digital versions.

Sony also suffered a disastrous server crash on a performance night. So HP was called in, and they called us. We modified our API to carry out closed captioning in real time. In fact, Meg Whitman, HP's CEO, demonstrated our system at the Mobile World Conference in 2016.

This opportunity, in turn, can lead into video or videoconference translation. We'd like to integrate our tech into YouTube, for instance, providing either captions or text-to-speech. Of course ASR for videos is quite challenging, what with multiple people speaking and overlapping, background noise, music, and so on. For greater accuracy, we match ASR results against our database.

One central point is that it's often easier to monetize human interpreting than automatic interpreting, so the latter can be used to upsell the former.

The question of automatic versus human interpreting is interesting. For instance, in our contract with Hewlett-Packard, we have five different contracts for different ranges of services. One of these is our IntelliConference conference calling. In this service, human and machine translation can be mixed for multilingual conversations.

Some of our customers use mostly human translation, and some mostly machine. In a customer hospital in Buffalo, for example, 200 of their clinicians have access to our hybrid system. They can use automatic translation, and many tell us that this is adequate for diagnosis, and they appreciate the cost savings. However, there is a failsafe button on their interface which invokes a live interpreter when necessary. Human and machine translation can be used to check up on each other. Even if you have a human interpreter, you may want to see the machine translation as verification. Or you may just want to see ASR transcription of the source language as a comprehension aid.

Language pairs: We do voice-to-voice translation for 44 languages. For text translation, it's about 80 languages. For either voice or text, any combination of languages is possible. We also offer human interpreting for 250 languages, reachable within one minute.

SWOT analysis:
Strengths: We have a low-cost, highly efficient, end-to-end translation solution on any platform, for speech or text, for any kind of content. And we always keep real-world needs in mind. We recently heard of a case in which a caller to 911 died because no translator was on hand. That's the kind of situation we're setting out to eliminate.

Weaknesses: We're a bootstrapped company: we don't have the resources of a Google. To compensate, we have to stay lean and agile. We have to keep evolving and innovating to keep one step ahead of the big guys. But in dealing with big customers, we need to build the level of trust over time, as we have done with Hewlett-Packard.

Opportunities: We see self service as the primary way to scale. If we make our technologies easy enough to consume, they can penetrate any industry. Because evaluation of translation is so subjective, you'll probably see self-driving cars on the road before you see the UN using automatic translation.

We have to continually analyze how customers use our offerings, and it can be surprising. In one case, we found that many customers of our speech translation app were actually getting more use out of the built-in currency converter.

We're especially interested in enabling the use of our technology to mix in with an unlimited number of applications – since no one can know in advance what all these applications will be. For example, users may want to give verbal orders to Salesforce, or they may use our translation tech to recruit overseas employees speaking other languages. We recently worked with Alcatel on sentiment analysis: they wanted to run keyword analytics on their systems and the millions of customers who buy their smart phones.

Threats: The big boys can and will come out with their own automation platforms. However, they have their own threats: they're in danger of becoming just large marketing companies.

John Frei, Co-founder & CEO

John Frei was born in Bronx, New York, in 1974. He earned his B.A. in Economics from Rutgers University in 1998. Frei spearheaded the development of a software solution to automate the Applications process, saving the organization over 500 labor hours yearly. Frei relocated to Toms River, NJ, after being named Land Development Manager of Ryan Homes in 2005, he was responsible for managing over 60 communities representing over $100 m in revenue through the Entitlements process. In 2008, Frei formed FREI, LLC a construction Renovation & remodeling company. The organization completed renovations on over $2.5 million of Real Estate prior to Liquidating. In 2009, Frei created a proof of concept and applied for Provisional Patent for the SpeechTrans Technology.

Yan Auerbach, Co-founder & COO

Yan Auerbach was born in Moldova on July 20, in 1986. He earned his B.S. in Computer Information Systems from Pace University in 2008. Auerbach is responsible for the Systems Architecture as well as daily development, operations, and innovations of the product. He is also involved with marketing and building business relationships. Yan has been emerged in IT since he assembled his first computer at 9 years old. Previously Yan worked for the oldest Venture Capital Firm, BVP as Systems Administrator. At Bessemer he was exposed to the latest technological innovations. Prior to Yan's work at BVP, he was at Fidelus Technologies, deploying VoIP products to law and CPA firms.

SpeechTrans Inc., USA, 2010
Number of employees: 14
http://speechtrans.com/
S2ST product: SpeechTrans Ultimate IOS, Android and API

Spoken Translation, Inc. (Mark Seligman)

Origins and motivation: On completing my Ph.D. in computational linguistics, I took a research position at ATR from 1992 to 1995. I arrived in Japan just in time to participate in the first international demonstration of speech-to-speech translation early in 1993, involving teams in Pittsburgh, Karlsruhe, and Takanohara, near Nara.

My first job was to enable German morphology as part of the Japanese-to-German speech translation path. From there it was a smorgasbord of related research topics: example-based translation; the use of pauses to segment utterances; the use of co-occurrence information for topic tracking; software architecture for speech translation; and two or three more.

Once back in the US, I proposed to CompuServe a "quick and dirty" (pipeline architecture) speech translation demonstration, based upon their experimental translating chat. The result was the first successful demonstration, in 1997 and 1998, of unrestricted – that is, broad coverage or "say anything" – speech-to-speech translation. Crucial to the success of the demo were the facilities for interactive correction of speech recognition errors.

Following up on this experience, I founded Spoken Translation Inc. in 2002, adding verification and correction of translation to the mix. We decided to address the healthcare market because the demand was clearest there; and, after a long gestation, we mounted a successful pilot project at Kaiser Permanente in San Francisco in 2011. Since then we've refined the product based upon lessons learned in this pilot. Going forward, we need to supply an API so that our verification, correction, and customization tools can be retrofitted to almost any translation or speech translation system.

Technology: Converser for Healthcare, our proof-of-concept product, was pilot tested at Kaiser Permanente in 2011. However, it's only one implementation of our core intellectual property – verification, correction, and customization technology. Other implementations will be needed in the future to fulfill the IP's potential.

In this first implementation, presently limited to English and Spanish, we use a rule-based MT engine supplied by a small Costa Rican firm called WordMagic. It was selected because its interactive orientation offered an easy path to implementation of our verification and correction tools.

For ASR, we're using Nuance's Dragon NaturallySpeaking software, channeled through the cloud by SpeechTrans, Inc. The recognition quality is quite good, and impressively resistant to noise. This arrangement does keep the system dependent on the Internet, but this limitation is tolerable in our use case. Happily, Nuance software is now fully speaker-independent, so even the short voice profiling needed in our 2011 pilot is no longer necessary for either language.

For TTS, too, we're using Nuance software – in a somewhat antiquated version. The quality is adequate, but a number of later refinements are lacking. For instance, question intonation isn't well rendered in this version. After we built it into Converser, Nuance purchased an Italian company, Loquendo, whose voices are clearly superior. We'll hope to use those, or comparably refined voices, in the future. We should also switch automatically between male and female voices when we know the speaker's gender.

Our present use of rule-based MT has good and bad points. On the good side, rule-based systems can be debugged much more systematically and explicitly than the presently dominant statistical systems. But of course the statistical systems are now dominant for many good reasons – scaling above all – and we need to demonstrate our adaptability to these systems soon. Our patents present a design for doing so. Meanwhile, the neural network paradigm is gaining steam very quickly, and research is needed on providing verification, correction, and customization tools for these systems, too.

Use case and market: We decided early on to address the healthcare market, since the demand for speech translation was clearest there. However, we began in the early 2000s, when the major components (ASR, MT, and

TTS) and the infrastructure (cloud computing, mobile platforms, the application market, etc.) were still immature and not yet ready to provide a really practical user experience. We were also naïve about the difficulties of penetrating very large and conservative organizations.

However, we judge that the time is right now on both of these fronts, provided that products can deliver the necessary degree of reliability and customizability. For us, reliability implies not only measurable accuracy but user confidence, and this in turn requires some form of verification and correction capability.

Customizability is especially important considering that a given use case like healthcare is not monolithic, but instead breaks down into many sub-use cases like nursing (which breaks down further into meal service, pain medications, intravenous use,...); pharmacy (prescription drop-off and pickup, consultation,...); eye care (informed consent for surgery,...); and many more.

And of course healthcare is not the only vertical market in which acceptably reliable and specialized products could prove profitable. We've demonstrated that our current techniques could also be used in law enforcement, emergency response, and the military. Our business model is still in flux. Presently our thinking runs to freemium-style marketing: offering the service free, perhaps for extended periods, in order to build the following, followed by enterprise-wide or office-wide licensing with maintenance fees.

Language pairs: In the US, the language most in demand for healthcare is clearly Spanish, and this is the only language handled in the current version of Converse for Healthcare. However, many other languages are also important for large organizations like Kaiser Permanente, and these largely depend on the local area. In San Francisco, for instance, there is a need for Chinese, Korean, Russian, Vietnamese, Tagalog, and a long tail of additional languages like Japanese.

This is why we've cooperated over the past few years with a company whose APIs cover more than 40 languages for speech and translation. The challenge still ahead of us is to integrate with such APIs in order to add our verification, correction, and customization tools to a broad range of languages and platforms. The APIs recently offered by Microsoft, Google, and other large companies are now possibilities; but these programs present themselves as black boxes, which complicates the necessary integration.

To ease the process, we could work with partners interested in developing speech and translation solutions specifically for cooperation with Converser. We're discussing this possibility with certain European members of our Technical Advisory Board. But for which languages? There's always a trade-off between quality and scaling. Google has massively scaled up the number of pairs that they can handle, but has avoided making any representations concerning quality.

SWOT analysis:
Strengths: Our verification and correction tools are our main contributions. They're intended to help move SLT systems past the threshold of reliability for serious vertical markets like healthcare, business, emergency response, etc. By reliability, we mean not only measurable accuracy but user trust, without which users just won't use the systems when errors could be harmful or embarrassing.

Customization is also important: users demand SLT systems that meet their special needs, so we've worked on ways to quickly add specialized translation memory: users should be able to browse and search memory for frequently used phrases, and memory should work seamlessly with full machine translation. Our concentration on verification, correction, and customization needs is unique at present.

Given this emphasis, we disagree with researchers who believe that SLT must still remain within specific domains to be effective. I'm proud of having organized in the '90s the first successful demos worldwide of unrestricted – that is, broad-coverage, say-anything-you-want – SLT. A key element of that success was the inclusion of interactive correction facilities. These days, the general-purpose, consumer-oriented SLT provided by Google, Microsoft, and others,

while obviously far from perfect, is certainly sufficient for many social purposes, and is playing an extremely important role by finally bringing speech translation to the masses after decades of research and promises.

Weaknesses: Our coverage of languages and *platforms* is still puny. As soon as possible, we should move beyond Spanish and the Windows desktop and adapt our tools to SMT and neural network translation and to many delivery platforms.

Opportunities: We believe users will pay for systems that are sufficiently reliable and customizable, thus breaking out of the commoditization trap arising from the general-purpose services now provided for free by the very large players. So far, our successful pilot project at Kaiser Permanente has at least shown that users in the healthcare setting – both patients and staff – can respond positively to SLT systems, even the early versions we used five years ago. The machine translation community as a whole is presently weak in its treatment of semantics. Since the advent of the SMT paradigm, semantics research has been neglected. Presently, only a few commercial systems offer deep semantic analysis, for example Fujitsu's ATLAS. However, there's now an opportunity to revive this area, e.g. by leveraging the ontologies ("knowledge graphs") now in use at Google, Microsoft, and IBM. Meanwhile, the burgeoning neural network paradigm can bring in perceptually grounded semantics, based on video and audio.

More generally, the opportunity is to view translation and natural language processing as specialties within the areas of machine learning and artificial intelligence. I'm a veteran of the AI boom of the '80s and a believer in "strong" AI in principle, but I doubt that human interpreters will be replaced anytime soon in use cases demanding specialized domain knowledge, common sense, or cultural competence. I think computers will breach these areas as the century unfolds, but don't hold your breath.

Threats: I've mentioned commoditization resulting from free SLT. This is the main threat to the field right now, though we can hope it's temporary. As compensation, the movement of SLT into the mainstream establishes its reality for potential investors. I've also mentioned one escape route that we see, via improved reliability and customizability.

Mark Seligman, Founder and President

Dr. Mark Seligman is founder, President, and CEO of Spoken Translation, Inc. His early research concerned automatic generation of multi-paragraph discourses, inheritance-based grammars, and automatic grammar induction. During the 1980's, he was the founding software trainer at IntelliCorp, Inc., a forefront developer of artificial intelligence programming tools. His research associations include ATR Institute International near Kyoto, where he studied numerous aspects of speech-to-speech translation; GETA (the Groupe d'Étude pour la Traduction Automatique) at the Université Joseph Fourier in Grenoble, France; and DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) in Saarbrücken, Germany. In the late '90s, he was Publications Manager at Inxight Software, Inc., commercializing linguistic and visualization programs developed at PARC.

In 1997 and 1998, in cooperation with CompuServe, Inc., he organized the first speech translation system demonstrating broad coverage with acceptable quality. He established Spoken Translation, Inc. in 2002. In 2011, STI's Converser for Healthcare product was successfully pilot tested at Kaiser Permanente's Medical Center in San Francisco. He now chairs the annual panel on speech translation for TAUS (the Translation Auotmation Users Society), and is preparing a report on the state of the art.

Spoken Translation, Inc., USA, 2002
Employees: self and numerous contractors
http://www.spokentranslation.com/
S2ST product: Converser for Healthcare

Another obstacle is the push toward a maximally seamless user experience via a maximally simple user interface. This push is certainly right in principle – we're all students of Steve Jobs and Johnny Ives in this respect – but on the other hand reliability must be enhanced, and this requires facilities for feedback and control which demand some attention, at least sometimes. We think a balance must be found, and have worked on ways for users to easily turn verification and correction on when reliability is most needed but off when speed is more important. Or the adjustment could be made automatically in response to programs' confidence scores.

Microsoft's testing has convinced them that users will reject feedback or correction tools; but we think that users will be more accepting in use cases where reliability is crucial, and that acceptance will also depend on the specific interface design.

There's a hope that machine learning, coupled with big data and perhaps some human curation, will yield SLT systems accurate enough to operate without feedback and correction even in critical use cases, but we see this hope as illusory. After all, even human interpreters typically spend some 20% of their time in clarification dialogues. Until such dialogues can be simulated effectively, some feedback will continue to be necessary.

Overall, this issue in SLT is a special case of a vital issue in AI: how to take full advantage of automaticity while keeping humans in the loop. The question arises in the context of self-driving cars, for instance. So we see our work on interactive monitoring as part of a larger picture. Another part is played by SLT interfaces that allow call-up of human interpreters when needed.

Translate Your World (Sue Reager)

Origins and motivation: In addition to developing software, I have the skills of a professional interpreter in several languages. As ASR became more generally usable from 2000, my company began to provide a suite of software that used various aspects of speech recognition, automatic translation, and human interpretation, tailored to specific clients' needs. The various systems are selectable per use case. On one hand, the software provides automatic speech-to-text and speech-to-speech translation (e.g. for cross-language training, meetings, etc). This facility provides communication where there was none in the past. Furthermore, the use of automation fosters inclusion for people who have been left out due to language or inability to hear.

On the other hand, when perfect communication is required, our software supports human simultaneous interpretation, delivered remotely online. In general, we have found that corporate clients take care to use full automation in the appropriate circumstances, and to use human interpretation for vital communications. Large corporate clients can now choose from a menu of software solutions, including ASR, MT, TTS, subtitles, human, etc. Our newest service is voice translation for telephone conversations. The various elements all function with WebEx, Skype, and related Internet-based communication solutions, and are expanding now to telephone and call centers. The output can be delivered via text, TTS, or human voice.

Technology: TYWI operates as a massive traffic controller. We don't supply engines for ASR, MT, or TTS. Rather, we gather the best software and/or human resources for a given client's use case. The software may include translation memory for glossaries or for scripted conversations (including pre-translated phrases for formulaic situations). TM is particularly useful for doctor-patient medical communications, customer service, and cross-language tech support. The most relevant technologies on the horizon for voice-to-voice translation are bots – which will play important parts in all speech technologies – and call centers that will go fully global, with one agent serving customers in dozens of languages.

Use case and market: We offer two pricing models for individual users and for API enterprise users: (1) First, we offer monthly subscriptions for individuals. These use an interface that pops up over other Web conferencing like Webex, Skype, Adobe Connect, etc. There's also a version of TYWI that provides its own conferencing, with sharing of screens, movies,

or cameras and built-in translation. (2) Second, we offer "by the minute" enterprise pricing.

These charges slide with actual usage of each service (text-to-text, voice-to-text, or voice-to-voice). With this pricing model, when employees text-chat to each other across languages, it's less expensive than using speech recognition and TTS to communicate across languages on the phone. The API model is complemented by an on-premises version, which permits corporations to install TYWI directly on their own servers for increased speed and security.

SWOT analysis:
Strengths: We're aiming to be the go-to full-service, one-stop shop meeting corporate needs anywhere in the world in the areas of text communication and speech translation. Via our services, business people can use the great language software of the world in the areas in which each excels. Combined, these solutions achieve excellent across-language communication. We do not white-label: instead, we advertise the work done by these fine language technologists.

Weaknesses: The obvious weakness arises because, until now, people have not needed to pronounce clearly, pace their speech, or speak factually to take advantage of today's amazing but imperfect automatic translation (MT). So part of our job is education: to teach the world how to talk to a machine, and how machines think during the translation process.

without requiring internal or external plugins per the World Wide Web Consortium (W3C). – Editor]

Threats: The introduction of automatic voice translation is much like that of computers. People in the translation and interpretation industry fear that their jobs will be taken over by technology and disappear. The opposite is true. Automatic voice translation will enable corporations to dramatically improve their relationships with their employees; enable universities to teach around the world; and enable businesses to go global.

This disruptive technology will catapult entrepreneurs into the global marketplace, create new kinds of businesses that are no longer confined to one city or state, and change the way we do business forever.

Sue Reager, President

Sue Reager is an executive, entrepreneur and inventor. Her inventions are licensed by Cisco Systems, Intel and others. Reager speaks 10 languages, and has lived and worked in 17 countries. She is columnist for Speech Technology Magazine and Examiner.com. As president of Translate Your World ("Tywi"), a linguistic software development company, Reager oversees the design and development of the Tywi software that enables people to converse in any language, teach and give presentations to global audiences, plus provide customer service in 78 languages. Reager is an industry expert in voice application translation, IVR localization and concatenation for audio, text, and web applications. Clients: UPS, Comcast, Bell Canada, Turner Broadcasting, General Mills, Caterpillar, Honeywell...

Translate Your World, USA, 2009
Number of employees: 20
http://translateyourworld.com/
S2ST product: Tywi

Future

As Yogi Berra said, it's hard to make predictions, especially about the future. Nevertheless, in this chapter, we'll try. Following are our best guesses about future directions for automatic speech-to-speech translation.

We'll proceed in three sections, from least to most speculative. We first look ahead at platforms and form factors; then at directions for improvement along current avenues; and finally at new technology, and in particular the nascent neural network paradigm.

Platforms and Form Factors

One obvious trend in speech translation's future is the technology's migration into increasingly mobile and convenient platforms. Most immediately, this means wearable and other hyper- mobile platforms.

Wrist-borne platforms integrating SLT are already available. SpeechTrans, for example, presently offers the tech via both a wristband and a smart watch. In the associated promotional video, COO Yan Auerbach plays a suitor who plans to propose to his beloved at a certain fountain in Central Park in New York City, and texts her to meet him there – in Russian, for this is his character's only language. Unfortunately, Yan can't find the fountain, and his questions in Russian to park denizens don't help. Fortunately, a stroller wearing the translating wristband approaches. He drives the device with voice commands: "Okay, SpeechTrans, translate from Russian to English" and "Switch" (to change translation directions). Success, and happy ending.

In the same spirit, translation apps from iTranslate and Microsoft are already available on iPhone and Android devices. Neither app appears to offer text-to-speech yet. In compensation, however, Microsoft's watch offering can exchange translations with a nearby smartphone.

Eyeglass-based delivery of real-time translation is likewise available now in early implementations. Google Glass did include the functions of Google Translate – but the device has now been withdrawn. Still, second- and third-generation smart glasses will soon appear; and, given the astonishing achievement of Google Translate's dynamic optical character recognition, it's inevitable that near-future smart glasses will support instant translation of signs and other written material in addition to speech translation. Meanwhile, SpeechTrans offers computational glasses incorporating translation functions.

And wristbands and glasses hardly exhaust the wearable possibilities. Startup Logbar is developing a translating necklace (as discussed in the interview above), and Waverly Labs will

offer earpieces with translation software built in. Further afield, the anticipated Internet of Things will embed myriad apps, certainly including speech translation, into everyday objects.

Not all such devices will succeed. The survivors of the shakedown will need to address several issues already discernible. First, some balance must be found between supplying all software on the local machine and exploiting the many advantages of cloud-based delivery. (Recently, Google has begun enabling users to download modules for individual languages to their local devices. As the computational and storage capacities of these devices grow, this option will increasingly be offered.)

Second, since special-purpose devices (necklaces, earbuds, rings) risk being left at home when needed, they'll have to compete with multi-purpose devices carried every day and everywhere (watches, smartphones, prescription glasses); so they may find niches by offering compelling advantages or by customizing for particular use cases, as in healthcare or police work. And finally, the new devices must be both nonintrusive and ... cool. Google Glass, while presaging the future, failed on both fronts; but Google and others, one can bet, are even now planning to do better.

Development of Current Trends

We'll now consider the further development of present trends in S2ST under these headings: improving statistical and neural machine translation; knowledge source integration; and the return of semantics.

Improving Learning-based MT

The heart of a speech translation system is its machine translation component, and learning-based translation methods – those of statistical machine translation (SMT) and neural machine translation (NMT) – are clearly dominant now with no end in sight. How can they be improved for the purposes of *spoken* language translation?

First, big data has undeniably played a crucial – if not decisive – role in improvement and scaling so far. More data is better data, and plenty more is on the way. The crucial new element for improving speech translation, however, will be increasingly massive *speech translation* data. To augment already massive text translation data – both parallel and monolingual – for MT training, along with already massive audio data for ASR training, organizations will be able to collect the audio (and video) components of natural conversations, broadcasts, etc. along with the resulting translations. Training based upon these will yield a virtuous circle: the speech translation systems built upon them will improve, and thus be used more, producing still more data, and so on.

Second, to massive raw data can be added massive correction data. Users can correct the preliminary speech recognition and translation results, and these corrections can be exploited by machine learning to further improve the systems. Google Translate has made a strong beginning in this direction by enabling users to suggest improvement of preliminary machine-made translations via the company's Translate Community.

And just as more data is better data, more corrections are better corrections. At present, however, there is a catch which limits the crowdsourcing community: to correct translations most effectively, users must have at least some knowledge of both the source and target language – a handicap especially for less-known languages. However, verification and correction techniques designed for monolinguals might greatly enlarge the feedback base. In any case, this exploitation of feedback to incrementally improve speech translation jibes with a general trend in machine learning toward continual learning, as opposed to batch learning based on static corpora.

A third way to improve SMT and NMT may be to integrate them with rule-based methods for some purposes – to create hybrid MT systems combining the best features of all methodologies. We've survived a paradigm shift from rule-based to statistical methods, and are now witnessing a virtual stampede toward neural methods (about which more below). The benefits of these revolutions are undeniable; but during such changeovers, earlier work is too often abandoned. The rule-based paradigm can't compete with its younger siblings in

terms of performance, scaling, or any aspect of independent learning. However, it still retains value from other viewpoints (for example, see [Boitet et al., 2010] for a debunking of several misconceptions concerning the advantages of SMT), and the countless person-hours invested in capturing linguistic knowledge within this paradigm should be leveraged rather than discarded.

Knowledge Source Integration

In our chapter on the past of S2ST, we characterized Germany's Verbmobil project as ahead of its time in attempting to integrate into speech translation systems such multiple knowledge sources as discourse analysis, prosody, topic tracking, and so on. And as noted, similar studies were carried out within the C-STAR consortium, likewise somewhat prematurely. The computational, networking, architectural, and theoretical resources necessary for such complex integration did not yet exist. These resources do exist now, though, as witness the success of IBM's Watson system[1]. Watson calls upon dozens of specialized programs to perform question-answering tasks like those required to beat the reigning human champions in Jeopardy, relying upon machine learning techniques for selecting the right expert for the job at hand.

In view of this progress, the time seems right for renewed attempts at massive knowledge source integration in the service of machine translation. Just a couple of examples:

- In many languages, questions are sometimes marked by prosodic cues only, yet these are ignored by most current speech translation systems. A specialized prosodic "expert" could handle such cases. Of course, prosody can signal many other aspects of language as well – pragmatic elements like sarcasm, emotional elements like anger or sadness, etc.
- As discussed above, the Microsoft translation system which powers Skype Translator features an auxiliary program that addresses issues of spontaneous speech by cleaning and simplifying raw speech input, for example by eliminating repetitive or corrected segments. This is the sort of auxiliary expert we're considering, though its integration is unusually straightforward, since it can be cleanly sequenced between speech recognition and translation.

While current techniques for knowledge source integration are already available for exploitation, the nascent neural paradigm in natural language processing promises interesting advantages: as we'll see below, neural networks appear especially attractive for such integration.

The Return of Semantics

Language is primarily a way of conveying meaning, and translation is primarily a way of assuring that, so far as possible, a surface structure segment conveys the same meaning in language B as in language A. In the face of this painfully obvious observation, it's striking how far present-day automatic translation systems have come without even slightly understanding what they're talking about. Philosopher John Searle has in fact notoriously observed that current translation programs, and more generally current computer programs of all sorts, function without true semantics [Searle, 1980].

Making an advantage out of this apparent drawback, Google research leader Peter Norvig and colleagues[2] observe "the unreasonable effectiveness of data": given enough data and effective programs for extracting patterns from it, many useful computational tasks – natural language processing among them – can be accomplished with no explicit representation of the task, and, in particular, no explicit representation of meaning. Translation has been perhaps the quintessential example.

We'll speculate here, however, that meaning is after all due for a comeback. We'll consider explicit (symbolic or vector-based) semantic representations first, and then touch upon implicit (neural network-based and other brain-inspired) semantics.

In considering explicit semantic representations, we should first of all review the difference between an explicit representation of the meaning of a language segment and a

1 See http://www.ibm.com/watson/.

2 See https://www.youtube.com/watch?v=yvDCzhbjYWs.

representation of its other aspects, for example, its syntactic structure.

Consider first the syntactic or structural analysis of a Japanese phrase on the left. In their original order, the English glosses of the relevant Japanese words would be "car, [object marker], driving, do, person" – that is, "car-driving person," "person who drives/is driving a car." The syntactic analysis shows that we're dealing with a noun phrase; that it's composed of a verb phrase on the left and a noun on the right; that the verb phrase is in turn composed of a post-positional phrase; and so on. This part-to-whole analysis makes no explicit claim about the meaning of the phrase, though this might be computed via programs not shown.

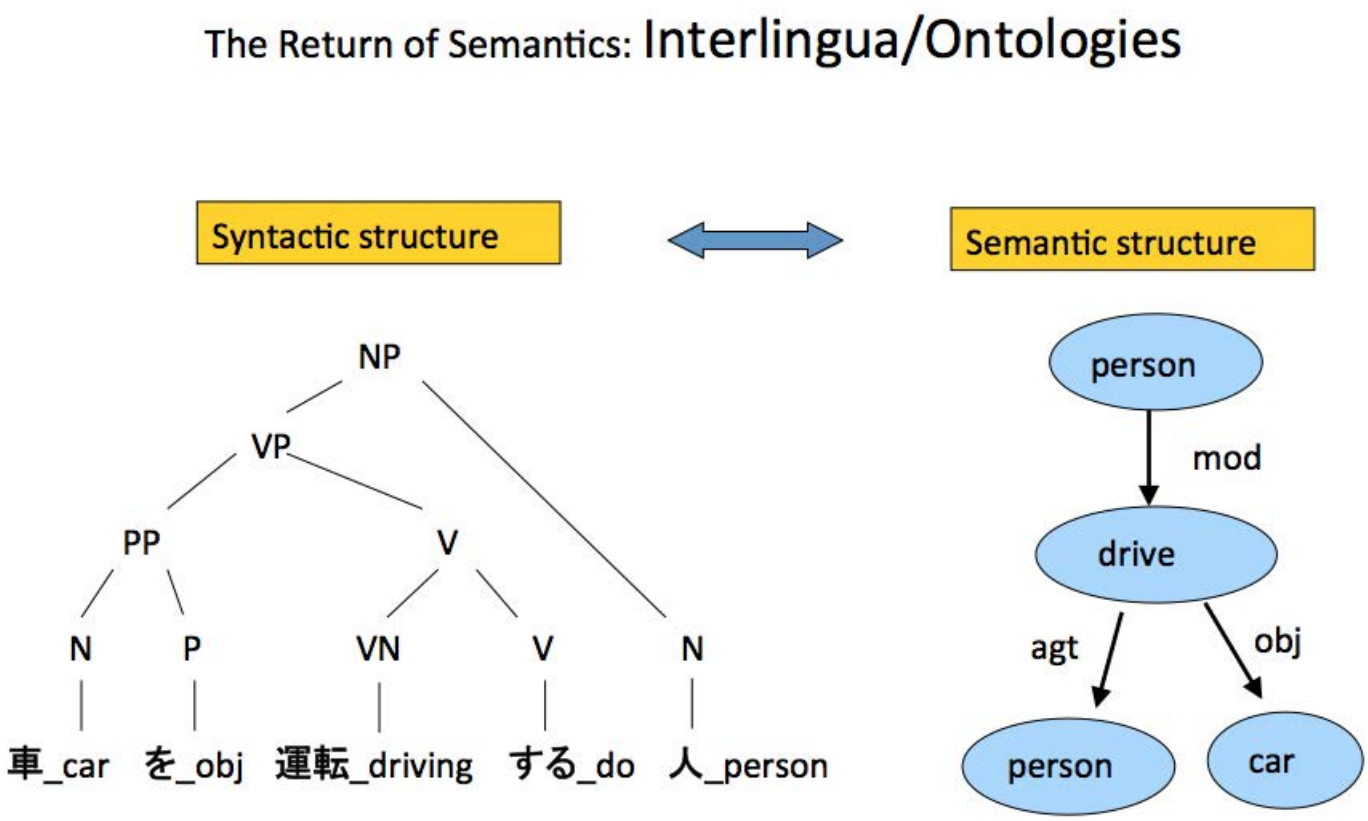
By contrast, on the right, we do see an attempt to capture the meaning of this phrase. PERSON is this time shown as a semantic object – presumably, one which could be related within an ontology to other semantic object such as ANIMAL, LIVING-THING, etc. The PERSON in question is modified – a semantic rather than syntactic relationship – by the action DRIVE, and that modifying action has an agent (the same PERSON, though the identity isn't shown) and an object, CAR.

Such meaning representations are far from unknown in natural language processing to date. On the contrary, there's a venerable tradition concerning their use. See, for example, Hutchins[3] concerning international research on interlingua-based MT, in which the goal of surface language analysis and the source of surface language generation was a representation aiming for maximal cross-linguistic universality. More recently, Hiroshi Uchida supervised creation of Fujitsu's interlingua-based ATLAS[4] system for English and Japanese, and then inspired international research based on the interlingua called United Nations Language, or UNL[5]. A dozen groups were involved worldwide.

However, the difficulties and frustrations of these efforts are well known. (See again Hutchins.) It's fair to say that explicit semantic representation for machine translation is now in the eclipse. Why then do we suggest that a comeback is to be expected, at least for some purposes?

In the first place, better results may sometimes be obtained, as illustrated by classic

3 At http://www.aymara.org/biblio/mtranslation.pdf.
4 See http://www.mt-archive.info/TMI-1988-Uchida.pdf.
5 See https://en.wikipedia.org/wiki/Universal_Networking_Language.



example-based translation cases: to render Japanese *kyouto no kaigi* or *toukyou no kaigi* fluently as "conference in Kyoto/Tokyo" or "Kyoto/Tokyo conference" rather than generically and clumsily as e.g. "conference of Kyoto/Tokyo," example-based systems exploited semantic symbols drawn from thesauri, indicating that *kyouto* and *toukyou* are examples of CITY and that *kaigi* is a type of MEETING[6].

Such processing advantages remain, though issues also persist concerning the costs and difficulties of obtaining semantically labeled corpora. It's also arguable that sufficiently big data can eventually yield such high accuracy via statistical or neural techniques that the advantages of explicit semantic processing for translation accuracy won't be worth the trouble.

However, other significant advantages of explicit semantics relate to universality and interoperability. Regarding universality, the original argument for interlingua-based MT was after all that the number of translation paths could be drastically – in fact, exponentially – reduced if a common pivot could be used for many languages. In that case, the meaning representation for English would be the same as that for Japanese or Swahili, and all analysis or generation programs could be designed to arrive at, or depart from, that same pivot point.

And concerning interoperability, the same representation could be shared not only by many languages but by many machine translation systems. Thus the ambition to overcome the Tower of Babel among human languages would be mirrored by the effort to overcome the current Babel of translation systems.

A common meaning representation, beyond bridging languages and MT systems, could also bridge natural language processing tasks. And in fact we do see explicit semantic representation taking hold now in tasks other than translation. Google, for example, has already begun to make extensive use of its Knowledge Graph ontology[7] in the service of *search*. "Thomas Jefferson" is now treated not only as a character string, but as a node in a taxonomy representing a PERSON, who is also PRESIDENT, who is also a LEADER, and so on.

This knowledge guides the search and enables more informative responses. Similarly, IBM's Watson system uses its own Knowledge Graph, this time in the service of *question answering* – initially focused especially upon the healthcare domain[8].

Unsurprisingly, Google and IBM presently use their own ontologies. However, eventual movement toward a common standard seems likely: one ring to rule them all, one semantic representation that could bridge languages, tasks, and competing or cooperating organizations. Meanwhile, efforts to inter-map or mediate among competing taxonomies also seem likely.

Before leaving the topic of explicit semantic representation, we should note that explicit doesn't necessarily imply handmade or non-statistical. One active area of semantic research aims to exploit the statistical relationships among words, or relationships between words and text segments of various sizes, to locate words in an abstract space, within which closeness represents similarity of meaning [Turney and Pantel, 2010]. (Intuitively, semantically similar words ought to participate in similar contexts and relations with other words.) The clustering in this space yields a hierarchy of similarity relations, comparable to that of a hand-written ontology.

Thus representation of a given word's meaning as a location in such a vector space can be viewed as an alternative to representation as a location in an ontology. A major advantage of the vector approach is its scalability: there's no need to build ontologies by hand. On the other hand, relations can be harder for humans to comprehend in the absence of appropriate visualization software tools.

Perceptually Grounded Semantics

We've discussed explicit (symbolic or vector-based) semantic representation, and now turn our attention to the more implicit, neurally inspired paradigm now taking shape. In

6 See https://en.wikipedia.org/wiki/Example-based_machine_translation.

7 See https://en.wikipedia.org/wiki/Knowledge_Graph.

8 See http://researcher.ibm.com/researcher/view.php?person=us-anshu.n.jain.

this direction, we'll consider the possibility of perceptually grounded semantics.

As a starting point, let's return to John Searle's claim that computer programs in general, and translation programs specifically, currently lack semantics. If shown a translation program making extensive use of explicit semantic representation as just discussed – that is, of symbols drawn from ontologies or of vectors drawn from vector spaces – he would be unlikely to change his mind. He'd probably argue that these semantic symbols or vectors, like the text strings that they accompany, could be manipulated blindly by a computer, or by a homunculus aping a computer, with no understanding at all of their content.

Such a translation system, he might say, could compute from the string "elephant" that an instance of the class ELEPHANT was represented, and thus, according to the rules, an instance of the classes PACHYDERMS, MAMMALS, ANIMALS, etc. (or a participant in the corresponding vector clusters). However, no matter how much further ontological, taxonomical, or relational information might be manipulated rule-wise, the system would still fail to recognize an elephant if confronted with one.

These arguments would be correct. The ontologically equipped system, despite its taxonomic sophistication, would remain devoid of any experience. But we can recognize now that this innocence is not an irremediable condition. It has now become possible for computational systems to learn categories based upon (artificial) perception. If, based on perceptual input, categories representing things like elephants are learned alongside categories representing linguistic symbols, and if associations between these categories are learned as well, then arguably the systems will come to have semantic knowledge worthy of the name – that is, perceptually grounded semantic knowledge which can eventually inform translation programs.

We can, for example, imagine a computational system that learns from examples to recognize members of the category CATS, thereby internalizing this category; learns from examples to recognize members of the graphic category NEKO-KANJIS (the Japanese character 猫, symbolizing the meaning "cat"), thereby internalizing this second category; and learns from examples to associate the two categories. We can also imagine a second computational system that learns likewise, but based on completely different examples.

And finally, we can imagine communication between the two systems mediated by transmission of newly generated instances of NEKO-KANJIS and confirmed through some objective functional test. The argument is then that, to both systems, instances of 猫猫 have a kind of meaning absent from handmade or vector-based ontology-based symbols divorced from (even artificial) perception.

This scenario could in fact be implemented using current technology. The DeepMind neural net technology acquired by Google can indeed form the category CATS (minus the label) based upon perceptual instances in videos. And as for the learning of communicative symbols like NEKO-KANJIS, in fact every speech recognition or handwriting recognition program already forms categories such that a new instance is recognized as belonging to the relevant category.

What remains is to learn the association between categories like CATS and NEKO-KANJIS, and then to demonstrate communication between computers whose respective learning has depended upon unrelated instances.

We can extend this story of perceptually grounded semantics to translation by assuming that, while one computational system learns an association with NEKO-KANJIS, the other learns a different linguistic symbol category instead, say for the written word "cat" in English. Then for communication to take place, a kanji instance must be replaced by an instance of that graphic element (or vice versa). If the replacement involved activation of a perceptually learned class in a third system, in which both the learned source and target language symbols were associated with the learned CATS category, then the translation process as well would be perceptually grounded.

Such demonstrations may emerge within a few years, but practical use of perceptually grounded semantic categories for automatic translation purposes will likely take longer, perhaps a decade or two. Meanwhile, integration of ontology-based and/or vector-based semantic representation might well move faster. The two strands of semantic research might thus proceed in parallel, with a top-down strand (explicit semantic representation) advancing alongside a bottom-up strand (perceptually grounded semantics). We can hope that the two strands will in time meet and enrich each other.

New Technology: The Neural Paradigm

As noted throughout, we're now witnessing a renaissance of the neural paradigm in natural language processing research and development, though many SMT and rule-based systems will certainly continue to operate. This rebirth (and yes, neural approaches to MT have been attempted before!) has been swift and dramatic. It took only a single year – 2016 – for the MT community to embrace NMT as dominant. Earlier neural modeling attempts couldn't quite outperform statistical systems; and then, last year, they suddenly and decisively could.

Work within the new paradigm has already had a major impact upon the state of the art in speech translation: Chris Wendt, head of the translation effort at Microsoft, has cited the advent of neural network-based speech recognition as a crucial breakthrough triggering the company's estimation that speech-to-speech translation was ready for prime time, and thus its decision that the time was right to release Skype Translator. (See Wendt's interview in the previous chapter.)

And now, just a few months prior to press time, Google has announced a major commitment to text translation based upon neural networks. Google's own announcement[9] summarizes the new development cogently, and a descriptive publication is available[10].

Viewed at a high level, current neural machine translation typically proceeds in two stages, comparable to the analysis and generation stages of traditional translation systems. In the first stage, the input is encoded as a list of vectors, in which each vector in the list represents the network's processing of all the input elements (words or characters) so far. When encoding of the input is complete, the decoder, or target language generator, begins term-by-term translation.

Each output element's translation depends on certain parts of the vector list created in the first stage; but exactly which parts depends on the network's prior learning concerning the relevance ("weight") of given source language elements to given target language elements. Thus the system appears to shift attention to the most relevant elements in the source list as each target language word or phrase emerges.

We mentioned superior performance in both speech recognition and translation as the most obvious trigger of the recent dramatic shift to neural methods. True, but this benefit is only one among many.

Yes, neural performance now appears to solidly exceed that of statistical approaches. For instance, Google claims that, as rated by bilingual humans, the new style reduces translation errors by more than 55%-85% on several language pairs when tested on sample sentences from Wikipedia and news Websites[11]. The team is putting the system into production for multiple languages, relying upon its proprietary machine learning tools and Tensor Processing Units for computational power. Thousands of language pairs are now supported, so there's considerable work ahead; and the announcement duly notes many substantive problems yet to be overcome. Even so, the entry of neural network translation into production use for text translation represents a milestone certain to influence the course of speech translation.
Neural models develop internal hidden representations of knowledge that can often generalize across tasks. Neural networks were born to learn abstractions. The "hidden"

9 See https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.

10 See https://arxiv.org/abs/1609.08144.

11 Sample neural translations can be compared with human and phrase-based statistical translations at https://research.googleblog.com/2016/09/a-neural-network-for-machine.html.

layers in a neural network, those which mediate between the input and output layers, are designed to gradually form abstractions at multiple levels by determining which combinations of input elements, and which combinations of combinations, are most significant in determining the appropriate output. The more hidden layers, the more levels of abstraction become possible; and this is why *deep* neural networks are better at abstracting than shallow ones. This advantage has been evident in theory for some time; but deep networks only became practical when computational processing capacity became sufficient to handle multiple hidden layers.

With respect to machine translation, this hidden learning raises the possibility of training neural translators to develop internal semantic representations *automatically and implicitly* [Woszczyna et al., 1998]. And in fact, if translation is trained over several languages, semantic representations may emerge that are independent of the languages used in training. Taken together, they would compose a neurally learned interlingua, a language-neutral representation of linguistic meaning. A successful interlingua could facilitate handling of under-resourced or long-tail languages, thus opening a path to truly universal translation at manageable development costs. Several teams have begun work in this direction [Le, 2016][12], and early results are already emerging: SYSTRAN, for instance, has already announced combined translation systems for romance languages[13].

Neural processing opens the way for *multimodal* training of translation and other natural language processing systems. The simulated neurons in a neural network aren't particular about the input elements they receive, since it's their job to learn which input elements are significant for the job at hand. (Thus implementers are relieved of many design decisions previously faced.) Words or characters are acceptable; but so are sounds or speech recognition lattices; and in fact, so are pictures, videos, etc.

For our purposes, the immediate promise is of closer coupling between speech recognition and translation. [Sperber et al., 2017], for example, observes significant translation improvements if NMT translators train on speech recognition lattices rather than only on parallel text. The neural systems learn to incorporate the similarities and probabilities of speech recognition hypotheses to improve translation decisions. Such training may contribute to greater robustness, more nuanced and situated translations, and overall to more integrated and useful interpreting systems.

Even more adventurously, Kyunghyun Cho of NYU and others have experimented with multimedia natural language processing involving photo input[14]. In a similar vein, we've already mentioned the possible use of neural networks to learn perceptually grounded deep semantic categories – for example, via input from videos or sensors – which could be exploited for translation: the technology that Google purchased from DeepMind might be put to use in this way[15].

To be sure, neural networks, deep or otherwise, are not the only participants in the dawning neural paradigm. For example, Ray Kurzweil has for several years been attempting, with Google's backing, the construction of artificial brains according to his Pattern Recognition Theory of Mind [Kurzweil, 2013]. Kurzweil's original impetus for this work was explicitly its application to natural language processing, though no results have yet been shown. In general, however, techniques inspired by neurons seem destined to play increasing roles in

12 See e.g. https://www.slideshare.net/eraser/googles-multilingual-neural-machine-translation-system-enabling-zeroshot-translation and http://www.kurzweilai.net/googles-new-multilingual-neural-machine-translation-system-can-translate-between-language-pairs-even-though-it-has-never-been-taught-to-do-so. See also http://www.aclweb.org/anthology/N16-1101.

13 See http://forum.opennmt.net/t/training-romance-multi-way-model/86.

14 See https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-with-gpus/.

15 Note that automatically learned internal representations in neural networks, such as the interlingua representations for translation purposes discussed just above, might or might not be perceptually grounded. In present research, they are not: they don't yet exploit visual, auditory, or other sensory input. They can nevertheless be viewed as embodying semantic elements in that they capture semantic commonalities in textual expression: they function as switching points or gateways toward target segments previously found by humans to have the same meanings as the source segments.

speech translation – and throughout natural language processing – over the coming years and decades.

As neurally inspired systems for linguistic and other processing gain prominence, an important issue comes to the fore: should we treat the systems as black boxes – as oracles whose only requirement is to give the right answers, however incomprehensibly they may do it? Or should we instead aim to build windows into artificial brains, so that we can follow and interrogate – and to some degree control – internal processes?

The black box path is tempting: it's the path of least resistance, and in any case brains have until now always been opaque – so much so that behaviorism ruled psychology for several decades on the strength of the argument that the innards were bound to remain opaque, so that analysis of input and output was the only respectable scientific method of analysis.

However, because artificial brains will be human creations, there is an unprecedented opportunity to build tracing, tracking, and control facilities into them. Thus it may be possible to open the black box to view and control, as they form, the familiar elements of symbolic artificial intelligence and natural language processing: instances, categories, relations, and the ontologies and semantic networks that they compose, or the reasoning relating to them. For example, the abstractions represented by connection patterns within networks can be seen as representing categories and sub-categories with probabilistic or fuzzy boundaries.

Along these lines, Google's recent work on multilingual NMT has exploited promising techniques for visualizing semantic groupings as vector-based clusters[16]. Alternatively, the connection patterns can be seen as stochastic rules amenable to chaining, such that the conclusions of one provide the premises of others.

The attempt to maintain traceability, comprehensibility, and a degree of control in speech translation and other NLP can be seen as part of a larger trend within artificial intelligence to keep humans in the loop. In Google's experiments with driverless cars, for instance, designers have struggled to maintain a balance between full autonomy on the car's part, on one hand, and enablement of driver intervention on the other. As we build fantastic machines to deconstruct the Tower of Babel, it would seem healthy to remember the Sorcerer's Apprentice: best to have our minions report back from time to time, and to provide them with a HALT button.

And by the way: humans can and should retain their places not only on the user side of S2ST, but on the provider side as well. As several of our interviewees reminded us, one important factor in the reliability of automatic speech translation is the option to opt out of it – to quickly get a warm body on the line when artificial interpretation is proving inadequate. In demanding use cases, facilities for quickly switching between or interleaving automatic and human interpreters will be essential, as will tools to aid those interpreters. We'll also need feedback tools to help users judge when a more human (and more expensive) intervention would be helpful.

16 See e.g. https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

References

Allen, Jonathan, Sharon Hunnicutt, Rolf Carlson, and Bjorn Granstrom. 1979. "MITalk: The 1979 MIT Text-to-Speech system." *The Journal of the Acoustical Society of America*, 65 (S1).

Alshawi, Hayan, David Carter, Steve Pulman, Manny Rayner, and Björn Gambäck. 1992. "English-Swedish Translation Dialogue Software." In *Translating and the Computer*, 14, pages 10-11. Aslib: London, November, 1992.

Boitet, Christian, Hervé Blanchon, Mark Seligman, and Valérie Bellynck. 2010. "MT on and for the Web." In *6th IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)* 2010. Beijing, China, August 21-23, 2010, pages 1-10.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics*, 19 (2), June 199, pages 263-311.

Cohen, Jordan. 2007. "The GALE project: A Description and an Update." In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Kyoto, Japan, December 9-13, 2007.

Eck, Matthias, Ian Lane, Y. Zhang, and Alex Waibel. 2010. "Jibbigo: Speech-to-Speech Translation on Mobile Devices." In *Spoken Technology Workshop (SLT), IEEE 2010*. Berkeley, CA, December 12-15, pages 165-166.

Ehsani, Farzad, Jim Kimzey, Elaine Zuber, Demitrios Master, and Karen Sudre. 2008. "Speech to Speech Translation for Nurse Patient Interaction." In *Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*. Manchester, England, August, 2008, pages 54-59.

Frandsen, Michael W., Susanne Z. Riehemann, and Kristin Precoda. 2008. "IraqComm and FlexTrans: A Speech Translation System and Flexible Framework." In *Innovations and Advances in Computer Sciences and Engineering*. Springer: Dordrecht, Heidelberg, London, New York, pages 527-532.

Frederking, Robert, Alexander Rudnicky, Christopher Hogan, and Kevin Lenzo. 2000. "Interactive Speech translation in the DIPLOMAT Project." *Machine Translation* (2000) 15: 27.

Fügen, Christian, Alex Waibel, and Muntsin Kolss. 2007. "Simultaneous Translation of Lectures and Speeches." *Machine Translation*

(2007) 21: 209.

Gao, Jiang, Jie Yang, Ying Zhang, and Alex Waibel. 2004. "Automatic Detection and Translation of Text from Natural Scenes." *IEEE Transactions on Image Processing* 13 (1), January 2004, pages 87-91.

Gao, Yuqing, Liang Gu , Bowen Zhou, Ruhi Sarikaya , Mohamed Afify, Hong-kwang Kuo, Wei-zhong Zhu, Yonggang Deng, Charles Prosser, Wei Zhang, and Laurent Besacier. 2006. "IBM MASTOR SYSTEM: Multilingual Automatic Speech-to-speech Translator." In *Proceedings of the First International Workshop on Medical Speech Translation, in conjunction with NAACL/HLT*. New York University, NYC, June 9, 2006.

Johnson, Melvin, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation." https://arxiv.org/abs/1611.04558.

Kumar, Rohit, Sanjika Hewavitharana, Nina Zinovieva, Matthew E Roy, and Edward Pattison-Gordon. 2015. "Error-Tolerant Speech-to-Speech Translation." In *Proceedings of MT Summit XV, Volume 1: MT Researchers' Track*. Miami, FL, October 30 - November 3, 2015, pages 229-239.

Kurzweil, Ray. 2013. *How to Create a Mind*. New York: Penguin Books.

Maier-Hein, Lena, Florian Metze, Tanja Schultz, and Alex Waibel. 2005. "Session Independent Non-audible Speech Recognition Using Surface Electromyography." In *Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, ASRU 2005. Cancun, Mexico, November 27- December 1, 2005.

Morimoto, Tsuyoshi, T. Takezawa, F. Yato, S. Sagayama, T. Tashiro, M. Nagata, and A. Kurematsu. 1993. "ATR's Speech Translation System: ASURA." In *EUROSPEECH-1993, the Third European Conference on Speech Communication and Technology*. Berlin, September 21-23, 1993, pages 1291-1294.

Le, Than-He, Jan Niehues, and Alex Waibel. 2016. "Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder." In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2016*. Seattle, WA, December 8-9, 2016.

Levin, Lori, Donna Gates, Alon Lavie, and Alex Waibel. 1998. "An Interlingua Based on Domain Actions for Machine Translation of Task-Oriented Dialogues." In *Proceedings of the Fifth International Conference on Spoken Language Processing, ICSLP-98*. Sydney, Australia, November 30 - December 4, 1998.

Och, Franz Josef and Hermann Ney. 2002. "Discriminative Training and Maximum Entropy Models for Statistical Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July, 2002, pages 295-302.

Olive, Joseph, Caitlin Christianson, and John McCary, eds. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer Science and Business Media: New York City, 2011.

Roe, David B., Pedro J. Moreno, Richard Sproat, Fernando C.N. Pereira, Michael D. Riley, and Alejandro Macaron. 1992. "A Spoken Language Translator for Restricted-domain Context-free Languages." *Speech Communication 11* (2-3), June, 1992, pages 311-319.

Searle, John. R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences 3* (3), pages 417-457.

Seligman, Mark, M. Suzuki, and Tsuyoshi Morimoto. 1993. "Semantic-Level Transfer in Japanese-German Speech Translation: Some Experiences." *Technical Report NLC93-13 of the Institute of Electronics, Information, and Communication Engineers (IEICE)*. May 21, 1993.

Seligman, Mark. 1996. "Interactive MT and

Speech Translation via the Internet." In *Proceedings of MIDDIM-96 (International Seminar on Multimodal Interactive Disambiguation)*. Col de Porte, France, August 11-15, 1996.

Seligman, Mark. 2000. "Nine Issues in Speech Translation." *Machine Translation*, 15 (1/2), Special Issue on Spoken Language Translation, June, 2000, pages 149-186.

Seligman, Mark and Mike Dillinger. 2011. "Real-time Multi-media Translation for Healthcare: a Usability Study." In *Proceedings of the 13th Machine Translation Summit*. Xiamen, China, September 19-23, 2011.

Seligman, Mark and Mike Dillinger. 2015. "Evaluation and Revision of a Speech Translation System for Healthcare." In *Proceedings of IWSLT2015*. Da Nang, Vietnam, December 3-4, 2015, pages 209-216.

Shieber, Stuart M. 2003. *An Introduction to Unification-based Approaches to Grammar*. Brookline, Massachusetts: Microtome Publishing. Reissue of Shieber, Stuart M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. Stanford, California: CSLI Publications.

Shimizu, Hiroaki, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. "Constructing a Speech Translation System Using Simultaneous Interpretation Data." In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT) 2013*. Heidelberg, Germany, December 5-6, 2013.

Sperber, Matthias, Graham Neubig, Jan Niehues, and Alex Waibel. 2017. "Neural Lattice-to-Sequence Models for Uncertain Inputs." In *Proceedings of the Association for Computational Linguistics (ACL) 2017*. Vancouver, Canada, July 30 - August 4, 2017. (under review)

Suhm, Bernhard, Brad Myers, and Alex Waibel. 1996a. "Interactive Recovery from Speech Recognition Errors in Speech User Interfaces." In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP) 1996*. Philadelphia, PA, October 3-6, 1996.

Suhm, Bernhard, Brad Myers, and Alex Waibel. 1996b. "Designing Interactive Error Recovery Methods for Speech Interfaces." In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) 1996, Workshop on Designing the User Interface for Speech Recognition Applications*. Vancouver, Canada, April 13-18, 1996.

Stallard, David, Rohit Prasad, Prem Natarajan, Fred Choi, Shirin Saleem, Ralf Meermeier, Kriste Krstovski, Shankar Ananthakrishnan, and Jacob Devlin. 2011. "The BBN TransTalk Speech-to-Speech Translation System." In *Speech and Language Technologies*, Ivo Ipsic (Ed.), InTech, DOI: 10.5772/19405. Available from: http://www.intechopen.com/books/speech-and-language-technologies/the-bbn-transtalk-speech-to-speech-translation-system.

Turney, Peter D. and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research 37* (2010), pages 141-188.

Waibel, Alex. 1987. "Phoneme Recognition Using Time-Delay Neural Networks." *Meeting of the Institute of Electrical, Information, and Communication Engineers (IEICE)*. Tokyo, Japan, December, 1987.

Waibel, Alex, T. Hanazawa, G. Hinton, and K. Shikano. 1987. "Phoneme Recognition Using Time-Delay Neural Networks." ATR Interpreting Telephony Research Laboratories. October 30, 1987.

Waibel, Alex, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. "JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies." In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1991*. Toronto, Canada, May 14-17, 1991.

Waibel, Alex. 1996. "Interactive Translation of Conversational Speech." *Computer 29* (7), July, 1996, pages 41-48.

Waibel, Alex, Alon Lavie, and Lori S. Levin. 1997. "JANUS: A System for Translation of Conversational Speech." *Künstliche Intelligenz* 11, pages 51-55.

Waibel, Alex. 2002. *Portable object identification and translation system*. US Patent 20030164819.

Waibel, Alex, Ahmed Badran , Alan W. Black, Robert Frederking, Donna Gates, Alon Lavie, Lori Levin, Kevin Lenzo, Laura Mayfield Tomokiyo, Jurgen Reichert, Tanja Schultz, Dorcas Wallace, Monika Woszczyna, and Jing Zhang. 2003. "Speechalator: Two-Way Speech-to-Speech Translation on a Consumer PDA." In *EUROSPEECH-2003, the Eighth European Conference on Speech Communication and Technology*. Geneva, Switzerland, September 1-4, 2003, pages 369-372.

Waibel, Alex and Ian R. Lane. 2012a. *System and methods for maintaining speech-to-speech translation in the field*. US Patent 8,204,739.

Waibel, Alex and Ian R. Lane. 2012b. *Enhanced speech-to-speech translation system and method for adding a new word*. US Patent 8,972,268.

Waibel, Alex, Christian Fügen. 2013. *Simultaneous translation of open domain lectures and speeches*. US Patent 8,504,351.

Waibel, Alex and Ian R. Lane. 2015. *Speech translation with back-channeling cues*. US Patent 9,070,363 B2.

Waibel, Naomi Aoki, Alex Waibel, Christian Fügen, and Kay Rottman. 2016. *Hybrid, offline/online speech translation system*. US Patent 9,430,465.

Wahlster, Wolfgang, ed. 2000. *Verbmobil: Foundations of Speech-To-Speech Translation*. Springer: Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.

Yang, Jie, Weiyi Yang, Matthias Denecke, and Alex Waibel. 1999. "Smart Sight: A Tourist Assistant System." In *The Third International Symposium on Wearable Computers (ISWC)* 1999, Digest of Papers. San Francisco, CA, October 18-19, 1999.

Yang, Jie, Jiang Gao, Ying Zhang, and Alex Waibel. 2001a. "Towards Automatic Sign Translation." In *Proceedings of the First Human Language Technology Conference (HLT) 2001*. San Diego, CA, March 18-21, 2001.

Yang, Jie, Jiang Gao, Ying Zhang, and Alex Waibel. 2001b. "An Automatic Sign Recognition and Translation System." In *Proceedings of the Workshop on Perceptual User Interfaces (PUI) 2001*. Orlando, FL, November 15-16, 2001.

Woszczyna, Monika, Matthew Broadhead, Donna Gates, Marsal Gavaldà, Alon Lavie, Lori Levin, and Alex Waibel. 1998. "A Modular Approach to Spoken Language Translation for Large Domains." In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA) 98*. Langhorne, PA, October 28-31, 1998.

Zhang, Jing, Xilin Chen, Jie Yang, and Alex Waibel. 2002a. "A PDA-based Sign Translator." In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI) 2002*. Pittsburgh, PA, October 14-16, 2002.

Zhang, Ying, Bing Zhao, Jie Yang, and Alex Waibel. 2002b. "Automatic Sign Translation." In *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP) 2002, Second INTERSPEECH Event*. Denver, CO, September 16-20, 2002.

Zhou, Bowen, Xiaodong Cui, Songfang Huang, Martin Cmejrek, Wei Zhang, Jian Xue, Jia Cui, Bing Xiang, Gregg Daggett, Upendra Chaudhari, Sameer Maskey, and Etienne Marcheret. 2013. "The IBM Speech-to-Speech Translation System for Smartphone: Improvements for Resource-constrained Tasks." *Computer Speech and Language* 27 (2), February, 2013, pages 592-618.

Appendix: Survey Results

83% have never used S2S technology before

43% believe that S2S business model should be licence, followed by pay-per-use (38%)

54% use S2S technology for travel, followed by online meetings (54%) and customer support (43%)

18% think the S2S business model should be a mix of wholly free and freemium

85% do not mind if the voice of the translator is very or somewhat different from the original voice

65% think that large companies such as Microsoft and Google will become the main providers of S2S

45% of the respondents think it will take about five years before S2ST will be widely used

93% of the respondents think S2S is an opportunity for the translation industry

About the Authors

Mark Seligman

Dr. Mark Seligman is founder, President, and CEO of Spoken Translation, Inc. His early research concerned automatic generation of multi-paragraph discourses, inheritance-based grammars, and automatic grammar induction. During the 1980's, he was the founding software trainer at IntelliCorp, Inc., a forefront developer of artificial intelligence programming tools. His research associations include ATR Institute International near Kyoto, where he studied numerous aspects of speech-to-speech translation; GETA (the Groupe d'Étude pour la Traduction Automatique) at the Université Joseph Fourier in Grenoble, France; and DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) in Saarbrücken, Germany. In the late '90s, he was Publications Manager at Inxight Software, Inc., commercializing linguistic and visualization programs developed at PARC.

In 1997 and 1998, in cooperation with CompuServe, Inc., he organized the first speech translation system demonstrating broad coverage with acceptable quality. He established Spoken Translation, Inc. in 2002. In 2011, STI's Converser for Healthcare product was successfully pilot tested at Kaiser Permanente's Medical Center in San Francisco. He now chairs the annual panel on speech translation for TAUS (the Translation Auotmation Users Society), and is preparing a report on the state of the art.

Alex Waibel

Dr. Alexander Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the Karlsruhe Institute of Technology, Germany. He is the director of the International Center for Advanced Communication Technologies (interACT). The Center works in a network with eight of the world's top research institutions. The Center's mission is to develop multimodal and multilingual human communication technologies that improve human-human and human-machine communication. Prof. Waibel's team developed and demonstrated the first speech translation systems in Europe&USA (1990/1991 (ICASSP'91)), the world's first simultaneous lecture translation system (2005), and Jibbigo, the world's first commercial speech translator on a phone (2009).

Dr. Waibel founded and served as chairmen of C-STAR, the Consortium for Speech Translation Advanced Research in 1991. Since then he directed and coordinated many research programs in speech, translation, multimodal interfaces and machine learning in the US, Europe and Asia. He served as director of EU-Bridge 2012-2015, a large scale European multi-site Integrated Project initiative aimed

at developing speech translation services for Europe. He also served as co-director of IMMI, a joint venture between KIT, CNRS & RWTH and as principal investigator of several US and European research programs on machine learning, speech translation and multimodal interfaces.

Dr. Waibel received many awards for pioneering work on multilingual speech communication and translation technology. He published extensively (>700 publications, >23,000 citations, h-index 78, (ref: google scholar)) in the field, and received/filed numerous patents.

During his career, Dr. Waibel founded and built 10 successful companies. Since 2007, Dr. Waibel and his team also deployed speech translation technologies for healthcare providers in humanitarian and disaster relief missions. Since 2012, his team also deployed the first simultaneous interpretation service for lectures at Universities and interpretation tools at the European Parliament.

Dr. Waibel received his BS, MS and PhD degrees at MIT and CMU, respectively.

Andrew Joscelyne

Andrew Joscelyne has been reporting on language technology in Europe for well over 20 years now. He has also been a market watcher for European Commission support programs devoted to mapping language technology progress and needs. Andrew has been especially interested in the changing translation industry, and began working with TAUS from its beginnings as a part of the communication team. Today he sees language technologies (and languages themselves) as a collection of silos – translation, spoken interaction, text analytics, semantics, NLP and so on. Tomorrow, these will converge and interpenetrate, releasing new energies and possibilities for human communication.

About TAUS

TAUS is a resource center for the global language and translation industries. Our mission is to enable better translation through innovation and automation.

We envision translation as a standard feature, a utility, similar to the internet, electricity and water. Translation available in all languages to all people in the world will push the evolution of human civilization to a much higher level of understanding, education and discovery.

We support all translation operators – translation buyers, language service providers, individual translators and government agencies – with a comprehensive suite of online services, software and knowledge that help them to grow and innovate their business. We extend the reach and growth of the translation industry through our execution with sharing translation data and quality evaluation metrics.

To find out how we translate our mission into services, please write to memberservices@taus.net to schedule an introductory call.