This document is part of the Network of Excellence "Multilingual Europe Technology Alliance (META-NET)",
co- funded by the 7th Framework Programme of the European Commission through the T4ME grant agreement no.: 249119.

A Network of Excellence forging the
Multilingual Europe Technology Alliance

Vision Document

Vision Group Interactive Systems:

Results of first two meetings

Editors: Joseph Mariani, Bernardo Magnini

Dissemination Level: Public

Date: 28 December 2010

Grant agreement no.	249119
Project acronym	T4ME Net (META-NET)
Project full title	Technologies for the Multilingual European Information Society
Funding scheme	Network of Excellence
Coordinator	Prof. Hans Uszkoreit (DFKI)
Start date, duration	1 February 2010, 36 months
Distribution	Public
Contractual date of delivery	n.a.
Actual date of delivery	n.a.
Deliverable number	n.a.
Deliverable title	Vision Document - Vision Group "Interactive Systems": Results of first two meetings
Type	Report
Status and version	Draft
Number of pages	25
Contributing partners	CNRS, FBK
WP leader	DFKI
Task leader	ILSP
Authors	Joseph Mariani, Bernardo Magnini, input from all members of the Vision Group "Interactive Systems" listed in Section 3.2., with Alex Waibel as Chair and Volker Steinbiss as Rapporteur
EC project officer	Hanna Klimek
The partners in META-NET are:	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
	Barcelona Media (BM), Spain
	Consiglio Nazionale Ricerche – Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR), Italy
	Institute for Language and Speech Processing, R.C. "Athena" (ILSP), Greece
	Charles University in Prague (CUP), Czech Republic
	Centre National de la Recherche Scientifique – Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (CNRS), France
	Universiteit Utrecht (UU), Netherlands
	Aalto University (AALTO), Finland
	Fondazione Bruno Kessler (FBK), Italy
	Dublin City University (DCU), Ireland
	Rheinisch-Westfälische Technische Hochschule Aachen (RWTH), Germany
	Jozef Stefan Institute (JSI), Slovenia
	Evaluations and Language Resources Distribution Agency (ELDA), France

For copies of reports, updates on project activities and other META-NET-related information, contact:

DFKI GmbH
META-NET
Dr. Georg Rehm
Alt-Moabit 91c
10559 Berlin, Germany

office@meta-net.eu
Phone: +49 (30) 3949-1833
Fax: +49 (30) 3949-1810

Copies of reports and other material can also be accessed via http://www.meta-net.eu

Table of Contents

1 Executive Summary

This report presents the first results of the consultations conducted within the META-NET "Interactive Systems" Vision Group, after a series of two meetings in a period of two months.
It has been mentioned that Interactive Systems based on Language Technologies (LT) are progressively being used in many different applications areas (such as telephony (and more generally mobile communication), Internet, car, aid to the handicapped, games)
This has been made possible as the background environment is now smart enough to accommodate, and actually need, such Interactive Technologies, and as the corresponding LT have reached an acceptable level of quality to make them efficient. This achievement in LT quality was made possible by the availability of large amounts of data, either created on purpose or obtained over the Internet, which allowed developing good LT engines based on statistical approaches. It was also greatly facilitated by the possibility to assess this quality through evaluation measures, and to estimate their adequacy with the needs, not necessarily error-free, of the targeted applications.
Many of the present Interactive Systems successful applications deal with the relationship between speech and text: text-to-speech synthesis, voice dictation (including emails and SMS), voice search over the internet, close captioning and translation of videos, audiovisual indexing and retrieval, speech translation through MT, as it is feasible to train engines to learn the relationship between a speech signal and its corresponding textual transcription. More should still be done in that area, such as meeting or teleconferencing transcriptions.
But a large research area is still opened regarding the availability of spoken or multimodal dialog systems, whereas the present systems in operation are still limited to very specific tasks. This would imply to provide the Interactive Systems with the capacity of understanding, of handling a dialog and of conducting a multiparty conversation, also taking into account prosodic information, and possibly non-verbal information coming from other modalities (eye gaze, or gestures, for example). To this regard, the Human Factors aspects should be carefully considered.
Just as for systems dealing with the speech to text relationship, the availability of training data is also crucial here, while the way to annotate this data is still a research problem and while the effort and cost of such a task appear very large. It would therefore be wise to consider in priority applications where the training data is not too difficult to obtain, or which could take advantage of the existence of tools in related areas (such as automatic transcription of talk shows on TV). Similarly, the existence of evaluation protocols for measuring the quality of dialog systems is a major issue, which still appears as a partly open research problem.

Multilingualism is a major challenge for the European Union, in order to allow its 501 million citizens to have access to information coded in their language and in any language at least of the other EU Member States. It could be the duty of the EC to ensure the availability of the Language Technologies, which would make it possible. The full range of the 23 official EU languages should be the target, as well as major regional languages such as Catalan or Basque. It could be part of the directive on information accessibility, presently primarily aimed at the handicapped, as the language barrier may appear as a handicap.
European languages should be considered as a public good and a common asset, which must be protected and promoted.
In order to achieve those visions, a complete research and innovation ecosystem should be put in place, from basic research to deployment in an international framework, through technology development, language resources (data, tools, evaluation and meta-resources) infrastructure, networking and public procurement. The corresponding programmatic instruments should benefit from adapted funding schemes, IPR rules and distribution mechanisms.
Such an achievement will require a strong political support from the European institutions and from the Member States, in a collaborative way.

- **Needs**
  - ***Domain-specific needs***
    - Need #1: Better core technologies
    - Need #2: Going to understanding
    - Need #3: Going to natural dialog
    - Need #4: Handling multilingualism
  - ***Domain-independent needs***
    - Need #5: More Basic Research
    - Need #6: Availability of Language Resources
    - Need #7: Availability of an Evaluation framework
    - Need #8: Consideration of the ethical dimension
    - Need #9: Appropriate programmatic instruments
    - Need #10: Research / Industry collaboration

- **Visions**
  - ***Domain-specific visions***
    - Vision #1. Interacting naturally with Agents and Robots
    - Vision #2. Communicating everywhere

- Vision #3. Technologies which help
- Vision #4. Bringing advanced interaction in social activities
- Vision #5. I speak your language!
- Vision #6. Gutenberg still alive
- Vision #7. My private teacher
- Vision #8. I know who you are

- ***Domain-independent visions***
  - Vision #9. Many languages, one Europe!

2 Introduction

2.1 The META-NET Vision Building Process

A central objective of META-NET is the preparation of a major concerted effort geared towards the creation of the needed technological foundation for the European multilingual information society. An essential instrument to this end is the forging of a strategic alliance involving, in addition to the top level R&D centres, the active participation of European LT and ICT industry and many private and public stakeholders, including the language communities themselves.

The Vision Groups are a central instrument within META-NET. Each of the three groups brings together researchers, developers, integrators and (actual or potential, corporate or professional) users of LT-based products, services and applications. The goal of the groups is to generate domain-specific visions and roadmaps in the form of technology forecasts. This includes ideas for innovative applications of language technology and scenarios for the future knowledge society which can be supported by advanced technology. The visions produced will be gathered by the Technology Council, which will consolidate them into a Strategic Research Agenda. The Agenda will contain high-level recommendations and suggestions for joint actions to be presented to the EC and national as well as regional bodies. The three Vision Groups are:

1. Translation and localization
2. Media and Information Services
3. Interactive Systems

The Vision Groups are scheduled to meet twice a year. In 2010, two rounds of meetings have been successfully completed. Their output will be discussed at the 1st Technology Council Meeting (16/11/2010), which, in turn, will prepare the draft SRA. Preliminary findings of the Vision Groups will be presented at META-FORUM 2010 (17/11/2010). Additional meetings will take place in 2011, the goal being to provide input to the SRA draft and to further elaborate the visions.

This document is intended to provide a distillation of ideas, opinions and visions expressed by members of the Vision Group Interactive Systems during its first two meetings. It aims at serving as a basis for discussion on the challenging and innovative LT based scenarios that need to be addressed by 2020 and providing seed ideas for the drafting of the SRA to the Technology Council.

2.2 Interactive Systems

We consider here Interactive Systems that use Language Technologies. Those systems are part of the wider field of Human-Computer Interaction (HCI), which comprises several communication modalities, including text and speech, but also vision, touch, gesture, or haptics. This field progressively moved from a bilateral communication between the human and the machine, to a collaborative activity involving the human and the machine in the accomplishment of a task, and now to the communication among humans through the mediation of the machine.

Just as other technologies involving language processing, progress was very slow reflecting the difficulty to handle the human language. Progress was made possible thanks to the use of automatic learning approaches, to the availability of large amounts of data to train the systems, to the existence of evaluation metrics and protocols to measure the state-of-the-art and progress and to the identification of applications which do not request a 100% accuracy. It's amazing to realize that the most difficult tasks, such as Pilot-Plane Dialog were considered initially, and that the most successful ones are presently those which can cope with speech recognition errors and imperfect ranking, such as voice search engines over the Internet.

This slow process resulted in the fact that many large companies quitted this field and that nowadays in Europe the area is mostly addressed by SMEs. The situation is different in the US, where large companies, such as Google, Microsoft, IBM, Apple, CISCO, or medium ones such as Nuance, are very active in that field.

3 Vision Group Interactive Systems

We looked for representatives from such European SMEs and for researchers in that field. Interestingly, we actually have in the VG several people who have both an academic background and an industrial one, whether they started in the former and ended up in the latter, or the reverse. Those people are of course very precious for providing a double vision of the field.

Although the development of such Interactive Systems was slow, the number of corresponding applications progressively expanded in many different areas, and they are now part of daily life for many people in applications such as Dialog in Video games, Customer care and technical support, (public) Information access (such as train time table) and transactions, Military applications (translation and training), Museum guides and public information kiosks, Car interfaces and navigation, Voice search, Speech translation on Smart Phones, eMail answering, Voice Dictation on laptop and Smart Phones, Call Centers, Telephone control or Aids to the handicapped.

Specifically, the sectors and players of this group, as well as representative technologies and applications are the following:

Fields: Telephone and mobile communication, Call centers, Internet navigation, Social Networks, Videoconferencing, Interpretation and translation, E-commerce, Finance, Healthcare, (Autonomous) Robotics, Car navigation, Security, Entertainment (Games), Edutainment, CALL (Computer Aided Language Learning), etc

Stakeholders: Telephones companies, Internet companies, Network companies (videoconferencing), Software companies, Translation companies, E-commercial companies, Banks, Robotics companies, Automotive industry, Edutainment and game companies, Audiovisual sector, Service providers, etc

Technologies: Speech recognition, synthesis and understanding, Spoken and Multimodal Dialog, Speaker and language recognition, Emotion analysis, Voice search, Information Retrieval (Question&Answer), Text analysis and synthesis, Topic identification, Speech Acts analysis, Summarization, Machine translation and speech translation, Sign Language Processing, Image and gesture analysis and synthesis, Computer graphics, Computer vision, Acoustics, etc

3.1 Recruitment process

A first list of external members was proposed by the T4ME partners, and a selection was conducted in order to ensure a good balance between industry and research, and among the various EU Members States. More names were then added to that list by the Standing Committee, and we also constituted a reserve list. Overall, 31 invitations were sent and we got 26 acceptations to participate in the VG and 5 refusals (mostly due to too much workload). 6 people agreed to participate but expressed that they were unable to attend any of the two planned meetings due to other prior commitments, given that the dates have been imposed, and 2 of them provided written contributions. Below we list the key figures concerning the recruitment process of the Interactive Systems VG (listing external members that are not in the META-NET consortium only):

- Initial member suggestions: 49
- Shortlist (invitations sent out): 31
- Confirmations: 26
- Refusals: 05
- Meeting attendants (cumulative): 19
- Participation through Written Contribution only: 02

3.2 Meetings

The Vision Group Interactive Systems met twice in 2010:
1st Meeting: September 10, 2010 at Aero-Club de France, Paris, France.
2nd Meeting: October 5, 2010 at Charles University, Prague, Czech Republic.
22 members participated in the Paris meeting and 16 in the Prague meeting. The participants in the VG and at the meetings are listed in the table below. In addition, 14 members provided written contributions overall.

	Name	Organisation	Country	Sector	1st meeting attendance	2nd meeting attendance
1	Axel Buendia	Spir.Ops	France	Robotics and Games	☒	☐
2	Aljoscha Burchardt	DFKI	Germany	NLP	☒	☒
3	Nick Campbell	Trinity College Dublin	Ireland	Speech Technology	☐	☐
4	Khalid Choukri	ELDA	France	Language Resources	☒	☒
5	Morena Danieli	Loquendo	Italy	Spoken dialogue, text analysis for text-to-speech and emotional text-to-speech	☐	☐
6	Gil Francopoulo	Tagmatica & IMMI	France	Natural Language Processing, standards	☒	☐
7	Simon Garrett	British Telecom	UK	eCommerce	☐	☐
8	Martine Garnier-Rizet	Vecsys & IMMI	France	Mobile applications, Call Centres, Language Resources	☒	☒
9	Edouard Geoffrois	DGA	France	Defence Applications	☒	☐
10	Joakim Gustafson	KTH	Sweden	Speech Technology	☒	☐
11	Jan Hajic	Charles University	Czech Republic	Natural Language Processing	☐	☒
12	Paul Heisterkamp	Daimler	Germany	Car industry	☒	☐
13	Mattias Heldner	KTH	Sweden	Speech Technology	☐	☒

14	Arjan van Hessen	Telecats & Twente University	Nether-lands	Research/PI Head of imagination Telecats	☒	☒
15	Timo Honkela	Aalto University	Finland	Speech Technology	☐	☐
16	Simon King	University of Edinburg	UK	Speech Technology	☒	☐
17	Jimmy (Siegfried) Kunzmann	European Media Laboratory GmbH	Germany	Human Machine Interfaces, Location based Service (LBS), Mobile Users	☒	☐
18	David van Leeuwen	TNO & Radboud University	Nether-lands	Speech Technology	☒	☒
19	Joseph Mariani	LIMSI-CNRS & IMMI	France	Speech Technology IS VG Co-Convenor	☒	☒
20	Bernardo Magnini	FBK	Italy	Natural Language Processing IS VG Co-Convenor	☒	☒
21	Bart Noe	Jabbla	Nether-lands	User Industry (educational soft-ware / handicapped)	☐	☐
22	Jan Odijk	Utrecht University	Nether-lands	Language Resources, MT	☒	☒
23	Mehmed Ozkan	Biomedical Inst. Bogazici Univ.	Turkey	Bio-Medical	☐	☒
24	Gabor Proszeky	Morphologic	Hungary	NLP, speech, MT	☒	☒
25	Steve Renals	Speech I/O	UK	Speech Technology	☐	☐
26	Giuseppe Riccardi	Univ. Trento (formerly AT&T)	Italy	Speech Technology	☒	☐
27	David Sadek	Institut Télécom (formerly Orange)	France	Telecommunications	☒	☒
28	Ruud Smeulders	RABO Bank	Nether-lands	Financial industry, User Industry (user of mobile services, call centres, translation services)	☒	☐
29	Juan Manuel Soto	Fonetic	Spain	Speech analytics, speech services	☐	☐
30	Volker Steinbiss	RWTH & Accipio	Germany	Speech Technology, Language Resources IS VG Rapporteur	☒	☒
31	Daniel Tapias	Sigma Technologies	Spain	Telecom Voice applications	☐	☒
32	Claire Waast	EDF	France	Automatized Call Centers	☒	☐
33	Alex Waibel	CMU & KIT and Jibbigo company	Germany & USA	Spoken Language Translation on mobile phone IS VG Chair	☒	☒

Table 1. List of VG Interactive Systems meeting participants

The first meeting was intended to achieve mutual understanding concerning the goals of both META-NET and the Vision Groups, to define the sector, collect needs and topics to be discussed in depth later on and discuss the appropriateness of the group's composition. No input was requested by the participants in advance to ensure the greatest possible open-mindedness. This first meeting also included the Media and Information Services Vision Group, with plenary parts and parallel sessions for the two VGs. After a plenary introduction of META-NET and of the Vision building process, the IS VG meeting was structured into the following four sessions:

Session 1: Mutual understanding
- Introduction of participants and discussion of the VG functioning (including the designation of the VG chair and rapporteur).
- Analysis of the scope of the field and sectors (What Interactive Systems cover)

Session 2: Irritation and Provocation
- Lessons from the past. SWOT analysis.

Session 3: The Vision
- Time horizon and scope of the visions
- Your visions

Session 4: Expected outcome and next steps
- Drafting of visions
- Additional names end expertise
- Preparation of next meeting

It was followed by a plenary wrap-up including the two VGs. Alex Waibel was designated as IS VG chair, and therefore chaired the meeting, starting at session 2. Volker Steinbiss was designated as Rapporteur. The outcome of the first IS VG meeting was presented at the ICT 2010 conference in Brussels on September 28, 2010 by Volker Steinbiss.

For the second meeting, written input by the Vision Group members was requested in advance, in the form of responses to the following five questions:

A. What are attractive, plausible, powerful, challenging, innovative language technology-based applications or combination of applications with Use Cases that in your opinion could be realized through massive concerted research, development, and innovation actions?
B. Can you think of novel research advances in Language Technology that would be needed to support real breakthroughs of type A?
C. What are expected technological, economic or social developments that have to be considered as prohibitive or enabling factors in the planning of A and B?

D. What are the instruments/programmatic tools/collaborative schemes that could be used in order to achieve A and B?
E. Please provide examples of LT-based (successful) Interactive Systems presently in use.

It was also mentioned that the aim could be to draw a matrix between:

1.) Grand challenges on problems/applications,

2.) Missing science/technology challenges to be researched;

The second meeting was structured into four sessions:

Session 1: Introduction, context, wrap up 1st meeting and working methodology

Session 2: Participants' Vision input (questions A-D above)

Session 3: Synthesis and priorisation of seed ideas for public discussion

Session4: Preparation of report

The meeting was chaired by Alex Waibel. The rapporteur Volker Steinbiss presented the report of the first meeting he gave at the ICT conference in Brussels.

At the meeting, twelve contributions were presented by the IS VG members.

3.3 Coverage

The coverage of the field, once it has been delimited, was one of the topics of the first meeting. The area of medical applications was identified as missing, and a new external participant was invited to join the VG, and accepted. It was also mentioned during this first meeting that the technologies scope should include written language, Gesture I/O, emotions/affect and multimodal communication, and that the social dimension of applications should be considered.

4 Visions on challenging and innovative LT-based scenarios

4.1 Domain-specific needs

Regarding the research and technology needs, there is a common agreement that there should be:

1. **Need #1: Better core technologies**

   This should be obtained through **more Basic Research**, as a generic issue (see 4.2). It is necessary to improve the quality of **Automatic Speech Recognition** and to lower the Word Error Rate, which is still too high for noisy environments, far-field microphone, open vocabulary, and for many languages, including dialects or accents. The performances vary a lot among the speakers, and should be more consistent. This improvement in

quality is necessary both for cooperative speech and for listening-in (3rd party observing human-human).

The same apply for **Automatic Speech Synthesis**, where it is needed to give speech synthesizers a control of the parameters with a linguistic or paralinguistic meaning that are crucial for conversational abilities (e.g. pitch, loudness and voice quality). In order to develop better complete systems, it is proposed to use joint models for speech synthesis, recognition, translation, etc., and to investigate the use of models for recognition and synthesis that include factorization for information like speaker, accent, vocal age, etc.

Progress is also needed in bi-modal recognition by using the combination of lips and tongue movement recognition to increase speech recognition accuracy, especially in noisy environment. The same apply for improving the quality of talking heads.

Gestual language is a very active research area and Sign Language Processing for providing accessibility to the deaf is showing rapid progress, which still needs more efforts both for analysis and generation.

More basic research is also needed regarding physiological modeling, in order to better understand the speaking mechanisms and allow for applications in the medical area, such as neuro-muscular prosthetic interfaces that could integrate with the existing speech technologies. It is needed to have a better understanding of selective listening and to conduct basic research on physiological modeling of speech production and speech perception.

2. **Need #2: Going to understanding**

Even if, as previously said, more research is needed for speech I/O, it has reached a performance level which allows to consider more advanced applications that now needs understanding, should it be on textual, OCR, spoken or gestual support.

This requires to take into account the context, with rapid contextual linguistic and domain knowledge modelling, and to adapt to new situations, new domains, as a portability issue. In order to do so, to be able to continually learn in an unsupervised manner, including learning from mistakes, to mimic human learning behaviour and leverage virtually unlimited language resources that may become available once privacy and storage hurdles have been overcome.

Understanding means also the need to take much more into account the prosody and visual cues. Those are mandatory for emotion detection and production (via voice and/or gestures), and includes the development of lexical models for laughter and other non-text-related vocalizations - and a syntax / ontology for modelling their use.

The move from appointing an answer to an incoming question based on statistics to an interpretation of the question (what does it mean) requires the collaboration of language and speech technology experts with psychologists and communication experts.

3. Need #3: Going to natural dialog

Better voice I/O and more understanding allow for considering human-machine dialogs, where voice plays a major role as the most efficient communication modality, whenever its use is permitted. A real breakthrough in the use of LT in human-machine dialogues can be achieved if we can bridge the gap between questions/remarks (written via a web application or e-mail and spoken via speech recognition) and understanding.

In order to do so, work should be conducted on dialog models. Dialog systems shouldn't be only reactive to the human stimuli, but pro-active, with the ability of initiating dialogs. They should be able to understand that a voice emission is in their intention, which probably means that they should be personalized and able to follow a conversation among humans.

It means constructing conversational speech model, possibly taking all available modalities into account, and utilizing the conversational behaviour of all interlocutors and relationships formed between them for various interaction communication situations ((telephonic) conversational speech, meetings, etc).

The human factors aspects should be carefully studied, as, for example, voice interaction is to be preferred when vision or hands are busy, but is not appropriate in public surroundings as it lacks privacy.

It implies the processing of the Speech Acts. How Speech Acts are correlated with each other? What kinds of sequences are generated? What are the possible answers to a specific Speech Act? It is necessary to evaluate the broad range of Speech Acts (while usually research investigations are focused on cooperative informative types), like lies or humour for examples. The study of the sequences of speech acts is huge, and it surely depends on cultural factor, so we would have to isolate those factors, and try to find where those factors come from (in the decisional process) and how they change the choice of the response. It could be some kind of public cooperative research.

Interaction should be natural, which necessitates the development of "transparent" systems: systems which enable natural spoken interaction, without the user being continually aware that they are talking to a computer, systems which work well in natural unconstrained environments: multiple microphones, multiparty conversations, many acoustic sources, noise (mostly non-stationary noises, background speech, echoes, etc.).

It also involves considering non-linguistic information and the development of non-speech sensor devices to further multimodal speech processing - RFID tags, non-invasive motion-capture, vad, GPS, orientation, proximity, etc – where the existence of iPhone-like devices can probably help.

Construction of such dialog systems could be incremental, and take advantage of the availability of conversational data on the Internet.

It would require research and development activities on the usability aspects of person machine communication and in new paradigms of human interfaces.

Collaboration between the academic world and service companies/organization (customer contact centres, public services, hospitals, etc.) can lead to an easy and fast collection of a huge amount of real world questions, eventually combined with answers given by human experts.

4. **Need #4: Handling multilingualism**

   Multilingualism is a major challenge, especially for Europe. Even if it concerns all LT and all LT-based applications, we would however like to stress that it is especially needed for Interactive Systems development, where the language barrier should be addressed in order to allow people to communicate whatever the language they speak and understand. This is needed for many of the applications envisioned.

   It implies that the systems developed should include features allowing for an easy portability to various languages, and at least the 23 EU official languages and the major regional languages, such as Luxembourgish, Catalan or Basque, and for cross-lingual abilities, such as speech translation and information retrieval from text or audio, in any language.

   It also needs to consider the cross-cultural dimension, which would require collaboration with socio-psychological researchers, from ethnologists (what are socially accepted interruption patterns in this cultural community?) up to and including film semiotics (how do we commonly display this or that kind of social interaction such that it is understood by audiences in most cultural communities).

The enabling or prohibiting factors (some generic) related to the realization of the needs are the following:

4.1.1 Prohibiting factors

4.1.1.1 Society and Economy

Cultural, political and economic. There is in some countries a cultural, political and/or economic lack of interest for the languages spoken in the country, or an ignorance of the existence of Language Technologies, and on how they may support the use of those languages. There is also a neglectance of local linguistic and cultural variation, not to mention of people with limited language abilities or strong accents, such as migrants.

Privacy and Ethics. Privacy appears as a major problem. It has to be addressed in order to avoid the lack of usable Language Data, given that voice and image carry information on the identity of a person. Calls may not be recorded for the use outside the companies or may be

stored and used for only a couple of months (and laws may differ in the various European countries). As for the acceptance of audio monitoring for voice activation in public places, one precondition is to ensure from the very beginning the strictest privacy of everything that is monitored (cf. current discussion of Google street view in Germany). On the other hand, public attitudes can change: acceptance of surveillance cameras (CCTV), e.g., is immensely higher in the UK than in virtually all other countries. Collecting personal data might be too intrusive. It is also feared that privacy laws and copyright laws will make potentially very useful language resources unavailable to the genuine researcher, while large corporations (telecom, Webservice operators) may obtain similar data _anyway_, and gain unfair advantages w.r.t. the academic researchers.

This may be extended to ethics in general. There is also a need for invasive clinical research, including human and animal experiments, which raises ethical issues, and brings difficulties of getting approvals.

Price for personalized systems. While research aims at providing personalized systems, development of mass services (telephony systems, telco carrier type) is rewarding but development of individualized systems running on a Smartphone might not be, due to strong price competition in software and apps market.

Regarding the medical applications, there is a relatively small market size (compared to telecommunications), and therefore less immediate commercial interest.

4.1.1.2 Technology and Science

Limited knowledge. In some application areas such as medical applications, there is a very limited understanding of cognitive processes that would lead to successful physiological models, and a limited scientific background in modelling neuro-muscular structures at cellular level.

Technological complexity. Generally speaking, dealing with language is a very hard problem, of unlimited complexity. Some participants of the VG believe that, technologically, we have to concentrate on parallelism. While there are more and more chips available, we unfortunately do not know what to do with them (not only the high-level programming tools but the potential applications for them are still missing). What is needed is a parallel description paradigm of linguistic abilities, which are interconnected to social and cultural background of humans.

Availability of Language Resources. The cost of data collections - transcription/annotation, is huge as it still needs manual intervention. Data exist in very small quantity for some languages. Language resources tend to be collected and owned by larger companies, and financial exploitation of public resources hampers research. It is amazing that extracting data from YouTube is possible but that scientific data collection still prohibited in

many cases.

4.1.2 Enabling factors

4.1.2.1 Society and Economy

Societal changes that might influence the need for conversational systems are:

Ageing. Demographic change might increase the need for assistive technologies, including human-system interaction, and the demand for systems that pro-actively provide guidance in every-day situations, as they can help to 'cover up' short-term memory glitches (straightforward example would be an apartment equipped to answer the question 'Where did I put my glasses?'). Demographic and economic pressures mean that home care and support systems will become commonplace; and such systems will benefit from personalized spoken interaction.

Globalization. With globalization, there is a need to allow people to have access to information encode in languages that they don't understand, and companies to address markets where different languages are spoken. The increase of movable populations brings the need of systems with conversational skills.

Automatization of society and more efficiency. People are now used with the availability of broadband access to information, and to permanent access through mobile communication. In general, the questions proposed are "hot" in the industry and governmental organizations. In the 24/7 economy people expect an answer by return: waiting till Monday morning when the service centre is open again, is no longer a desired option! So systems that may facilitate automatic and intelligent answering may be considered as very useful both by the organizations/industry and the public. At the same time, the avalanche of information becomes more and more difficult to handle and needs the help of new technologies which rely on automatic text and speech processing.

Reduced costs. At the same time, equipment costs are coming down, enlarging the number of users while needing cheap and easy to maintain interfaces. Obviously, microphones, loudspeakers and cameras fall in that category.

Huge market. There is a potentially huge, while quite sparse market (entertainment, consumer apps, robotics), that would be opened up by the availability of reliable language technologies, such as an adaptive, controllable, understandable, expressive speech synthesis.

Green technologies. In order to reduce the production of CO2 which is claimed to be responsible for the climate change travels by plane, train or car should be made less frequent. Therefore remote meetings are becoming standard, stimulated by the economic conditions and climate change. In this framework, language technologies will enable much richer interactions in advanced videoconferencing services.

4.1.2.2 Technology and Science

Technological progress is also an enabling factor for the deployment of Language Technologies

Technology advances. We get the support of more computing power, storage and bandwidth, and we get over time a better understanding of the technology, and of the technology progress.

Ubiquitous technology availability (at low cost). We especially notice rapid developments in mobile computing - decreasing power consumption, high network bandwidth and cloud computing – which are stimulating demands for new interfaces. Mobile devices like Smartphones are networked and needs the availability of user-friendly ubiquitous technology solutions. Communities are trying out new technology and applications. Increased use of relatively small mobile devices with increasing computational power will be a push for developing more natural interfaces to such devices.

User-centric and crowd-sourcing. Today, the end-user is more in the loop. The development of successful language applications could make use of the users themselves (think web2.0). This could be done by providing non-experts with easy-to-use tools for building complete spoken dialogue systems, as well as to record and build new speech synthesis voices and to develop speech recognition for new languages and dialects, for example. New mechanisms to involve large numbers of non-experts now exist, such as crowd-sourcing (e.g. Amazon Mechanical Turk), and human computation team up with machine computation. Collaborative work conducted on volunteering bases is now common (see Wikipedia or TED translation, for example).

LR availability. A lot of Language Resources, such as corpus (text and speech) are now available, especially over the Internet. As data access becomes more open, the volume of available audio data will increase exponentially; flexible speech transcription will result in making such data oceans becoming searchable and structured.

4.2 Domain-independent needs

Regarding the research and technology needs, there is a common agreement that there should be:

5. **Need #5: More Basic Research**

   Some participants think that completely new approaches to Speech recognition, Speech synthesis and (Spoken) Machine Translation are required to realize the applications which are envisioned. Other think that statistical models that have been developed and that we want to use everywhere are not enough, partly because of the well-known sparse data problem, partly because of the nature of some human languages which do not really fit the paradigm. On the other hand, language technology cannot use many findings of

theoretical linguistics, because their descriptions are formalized for humans. Cultural background will also be needed to investigate: how can it be described and combined with linguistic abilities. E.g. speech-to-speech translation is not enough, but "situational" translation which would also be sensitive to the actual situation where it is used in.

6. **Need #6: Availability of Language Resources (data, tools, services and meta-resources)**

   Language Resources cover data, tools, services (such as evaluation) and meta-resources (such as standards, metadata, guidelines).

   Data. There is a need for more data to improve spoken language models, acoustic models, conversational models etc. and this for all the targeted languages. Addressing language understanding means the production of large amounts of data with various levels of annotation, including semantic annotation, the cost of which is huge, and even huger in a multilingual environment. The need is for real data, not artificially generated data. It is stressed that it would be needed to constitute a major corpus collection. We don't yet have unstructured conversational data but we will face problems re: privacy and other ethical issues that have to be solved first.

   The situation for low-resourced languages is especially crucial. Data should be produced for those languages, but other approaches could ne conducted in parallel, such as How to adapt models with less data (going from one topic to another, or from one language to another)? How to deal with low resource languages without much manual intervention, taking advantage of the existence of comparable data in other languages (especially those of the same family)?

   Tools. The availability of Open Source systems allows for lowering the entrance barrier for the community of research, technology and development entities. There is a specific interest for "parallel tools". The development of such tools should also be fully supported. Less and less end-users want to pay, and the future may mainly rely on the public support of research activities and free tools development, the output of which could be used by everyone.

   Meta-resources. Standards should be created and made available to validate, reuse and share language resources (data and tools). Language Resources should be easy to get, especially when their development was fully supported by public funds. There should be a clear policy regarding resources sharing.

7. **Need #7: Availability of an Evaluation framework**

   There is a very strong request for providing evaluation frameworks, in the different areas

of research and especially for the technologies that are marked as of key importance to the EU. This includes data collection, evaluation infrastructure, running evaluation, dissemination (workshops), research in evaluation methodologies and measures. It would be crucial to measure advances and identify new challenges, while comparing the Technological Readiness Level to the needs of the envisioned applications. Such evaluation should be as open as possible (availability of data and open participation). Given its infrastructural nature, it should be fully funded.

8. **Need #8: Consideration of the ethical dimension**

   This includes the privacy issue, which, as already mentioned, increases the problem of the production of language data for developing the systems. It is especially true for speech and audiovisual information, where the identity of people can easily be found through their voice or face.

   Privacy issues and data IPR should therefore be addressed as a major topic for allowing the development of reliable technologies, and all kind of non-disclosure guarantees will be needed. Anonymizing technologies should be aimed at, that can assure that only information relevant to the technology of interest is retained while other, privacy-bearing information is removed.

9. **Need #9: Appropriate programmatic instruments**

   Larger, longer projects. The sentiment expressed by the majority is that there should be place for larger projects, involving more efforts and a larger budget, lasting longer (5 years), with yearly evaluation and possible introduction of new partners following the same rules along the years, while one participant still proposes several IPs and STREPs, but coordinated in order to solve different research problems and to build components of the overall system (through research and industrial partners).

   Networking. While the need for a stable organization that coordinates the long-term endeavour is mentioned, the necessity of networking is also expressed.

   Grand Challenges. The preference goes to identify Grand Challenges that address specific technological issues, and will allow teams of similar background in Europe to collaborate to tackle those issues, rather than Integrated Projects where wide interdisciplinary teams have to collaborate and spend their time trying to understand the other partner's fields and/or built demonstrators that try to integrate subsystems that were not designed for that by the researchers in the first place: companies can do integration themselves, when the technology is ready and the market is there.

   It is important that any funding programs be goal-oriented and focus on the desired outcomes (e.g. speech recognition in unconstrained spaces with many acoustic sources) rather than focusing on the approaches to be deployed (e.g. it does not help to add constraints that things should be "cognitive" or "bio-inspired" or involve "new paradigms",

as these constraints tend to distort the field).

Appropriate funding. It has been mentioned that there is a big barrier with the need to come up with a 50% local funding for projects which do not imply return of investments, such as evaluation or Language Resource production.

End to End Research & Innovation ecosystem. While technological progress is aimed at, the need for keeping on conducting Basic Research is also stressed, as a lot of fundamental research is still required, so must be possible, and should be supported through government policies and incentives.

On the other extreme, the need for support to international deployment in order to help SMEs and research centres to launch novel technologies in the international market is also mentioned

10. **Need #10. Research / Industry collaboration**

Multidisciplinary. An effort like audio monitoring of social interaction must bring together knowledge from a wide variety of scientific disciplines. Many of those are in their thinking and in their attitude light-years away from contributing to an engineering effort having as a goal working technical solutions. LT researchers and industrials should try and come up with proposals on how humanities & social scientists (such as psychologist, sociologists, and philosophers) could deliver their contribution, preferably within co-operative projects.

Real World Data within a joint industry / public research collaboration. Strong collaboration between industry and academic institutions for the collection of real world data while guaranteeing the privacy of these data will be a key issue.

User in the loop. Collaborative-based applications should involve users in the loop.

Simple IPR framework. The IPR issue within EU projects should be made easier by adopting one or two common IPR schemes set. Examples could be:

- 100% funding of research/development. Resources and software will be GPL. EU supports distribution infrastructure.
- 100% funding of research/development. IPR will be with some EU institute, rules and conditions are known beforehand
- partial funding. Partners must resolve IPR rules before submission of proposal. EU has template for IPR conditions/procedures.

4.3 Domain-specific visions

1. **Vision #1. Interacting naturally with Agents and Robots**

Interaction with Conversational Agents (in games, entertainment, education, communication, etc). The development of more and more sophisticated video games calls for advance interaction, which comprise spoken dialog, like a chatting head or a conversa-

tional agent that you could use in any entertainment application, like video games, but also in education (e-Learning) and training, including serious games.

In those areas, but also more generally for communicating over the internet, self-learning context-aware personalized agent could be developed, gathering functionalities such as assistant, agent, self-learning (in particular adaptive), with speech, language and multi-modal input abilities, and speech, text and multi-media (e.g. icons on a map) output abilities, augmented reality (e.g. text / icons displayed in glasses), and covering human–machine, but also human–human mediated by machine (e.g. finding what person X said about topic Y) communication. It would include "simpler" tasks, such as processing of emails, sms's or telephone calls.

Interaction with robots. Such agents may be materialized as autonomous robots which would need natural interaction with humans, while they could learn by acting and from interaction.

Spoken dialog, also in instrumented spaces. It is now time to (re)consider spoken dialog, or conversational speech technology, in a natural way, showing advanced features: voice activated, no turn-key, cross-lingual (language adaptive), presence detection (context), position detection, using distance mike and loudspeaker, gaze tracking, target identification (multimodal), language identification, emotion detection and recognition, proactive dialog, social modelling, dialog management, 'always-on' microphone in an immersive environment. Such spoken dialog systems should be easily portable to new applications. Instrumented spaces (real and virtual) can use speech and language technology to support interactions and creative (in the widest sense) processes.

2. **Vision #2. Communicating everywhere**

   Mobile applications. Advanced mobile applications taking advantage of LT based Interactive Systems (beyond simple commands) should ne aimed at. "Mobile always on" applications could act as a personal assistant, learning through time without supervision.

   Augmented Reality. This could be couples with Augmented Reality, within applications aiming at "Feeling at home everywhere", in the form of multilingual and context aware mobile devices, featuring automatic translation (street name, menu, maps, foreign newspaper, radio & TV Broadcast…), and the description of the close environment as Augmented Reality in a multimodal way ("I take a picture and I ask a question regarding that picture."), also taking advantage of the existence of RFID tags.

3. **Vision #3. Technologies which help**

   Assistive applications. Interactive systems would be very useful for clinical and assistive applications, such as:

   - personalized speech technology systems for people with, e.g.,motor control problems,
   - home care interfaces for older people, or people with disabilities.

It would be needed (law on accessibility) to support tools for providing access to information for the handicapped (Blind, Hearing and/or Speech Disabled, Motion Disabled.) and Educational and rehabilitative tools for training the disabled.

In this framework, results of ambient intelligence developments will be combined with language technology applications. Meta-communicative elements of communication will also be taken into consideration (facial reactions, movements, gestures, etc, that is, more and more analog processes versus today's digital ones).

Sign Language processing. Sign language recognition, synthesis and translation would allow sign language speakers to communicate either over communication channels or with non sign language speakers in person to person conversations. This will allow deaf people to access to the same services than the rest of citizens and in similar conditions (with no need for a sign language interpreter).

4. **Vision #4. Bringing advanced interaction in social activities**

   Social applications. Social computing refers to the computational support for social interaction between people, and for the set of technologies which take into account social context. Social computing has attracted significant application interest from web entrepreneurs, game developers, marketing, PR, opinion analysis/polling, and practitioners of e-government. Natural human communication is at the heart of social computing and thus the intense activity in the area has much to gain from speech and language technologies. In particular applications which consider social context require a much richer loop between synthesis and recognition, are able to generate and interpret socially-plausible speech and act conversationally in a socially acceptable manner. Social applications should therefore include interactive devices which behave within social norms, which close the loop between recognition and synthesis, reacting to the conversation and the environment, conversational devices which do not demand 100% attention and can join a conversation appropriately, devices which adapt to different social settings and can both recognize and synthesize socially plausible speech (and not just extremes of emotion). Such systems should be able to work with large groups of people (e.g. during a workshop coffee break that takes place in an instrumented space) and to infer social interactions. This implies the modelling (and technology development) of such casual conversational interaction -social talk, rather than task-directed (text-based) talk - finding out how people use talk and spoken interaction to interact with each other at a social/ethological level. Social applications based on human-computer conversations might include domestic services (smart homes), games, and information gathering programs. Social applications based on human-human conversations, will include meeting support systems, richer videoconference / remote meeting systems, and companion devices able to keep track of the current communication context.

5. **Vision #5. I speak your language!**

Speech-to-Speech Translation. Online speech-to-speech translation (computer in the loop) should be made available in order to break the language barrier in communication. It should reach sufficient quality for conversational speech, including handling new words, and allow everyone to speak up his/her own language, in a natural way.

Interpretation in Videoconferencing. Such technologies could be used for real-time interpretation in video/tele-conferencing systems.

Cross-lingual information access. It would also provide access to the whole world from home within a multilingual WWW, including 200 to 1000 languages: Cross-lingual queries, Question-Answering, search in Natural Language (I'm looking for a photo where...), Automatic subtitling or even dubbing of (radio, TV, podcasts), Multilingual description of images and scenes (for visually impaired people), Automatic translation of textual documents, Multilingual speech synthesis, Query by Sign Language (webcam), Social networks with automatic translation (chats, twits, emails), Conversational agent.

6. **Vision #6. Gutenberg still alive**

   Going from speech to text, and from text to speech, with sufficient quality still allows for innovative applications.

   Speech transcription. It would be possible to provide fully automatic transcription services, for application in police interrogations, political debates, court cases, low-cost close captioning of audiovisual media, meetings i.e., anywhere where such transcription is carried out manually. Technology is only of added value if the entire event is recorded (audio, video) and indexed and made accessible by all relevant parties. It implies ASR, speaker diarization, language recognition, Low latency speech-to-text translation (for all European languages),.

   Videoconferencing. Advanced videoconferencing services could include the automatic transliteration of conversations.

   Close-captioning. In many EU countries, the audiovisual law will oblige private and public TV channels to subtitle 90% of all the TV programs in 2013. Most of the movies have already subtitles but there are live TV programs which should be subtitled in real time. The cost of doing it by humans makes it unviable. There is therefore a need for developing the speech recognition technology allowing automatic subtitling, while this technology can also be used for indexing audio and audiovisual sources of information for information retrieval.

7. **Vision #7. My private teacher**

   Interactive systems are necessary for the present development of new educative approaches which allow the teachers and the students to be in remote places. Courses can be automatically transcribed and even translated in the language of the student, together with the corresponding slides.

The system may itself ask questions, get answers from the student and assess the adequacy of the answers. It can be used for learning a language, also checking the quality of the spoken utterances (pronunciation, timber, intonation, rhythm, accent, etc) (CALL: Computer Aided Language Learning).

8. **Vision #8. I know who you are**

   Biometrics. The development and integration of multi-biometrics in interactive systems would allow identifying or authenticating users in different kinds of applications: From loading the most appropriate acoustic models (the user models) to granting access to information or allow introducing, update private information, allowing e-commerce, defence and security applications, etc. This implies research and development in detecting and solving vulnerabilities of biometric systems.

4.4 Domain-independent visions

9. **Vision #9. Many languages, one Europe!**

   A strong focus should be put on multilingualism, as it appears as a challenge for the EU. Instead of addressing 2-3 languages, all projects should take into consideration all the 23 official EU-languages, plus some major national or regional languages, such as Luxembourgish, Catalan or Basque. Provision should be made in order to facilitate the portability of any LT developed to the full scope of languages, taking into account the peculiarities of the various languages and language families.

5 Conclusions

The ideas developed within the two first months of the "Interactive Systems" Vision Group activity and reported here should now be discussed with our peers and with the various stakeholders, merged with the needs and visions expressed by the two other Vision Groups on "Translation and Localization" and "Media and Information Services" and organized in order to constitute a Strategic Research Agenda for the next decade.

In the conduct of this exercise, we must take in consideration the major challenge that multilingualism represents for the European culture and economy.