Supported by ELSNET

The Workshop Programme

Sunday, June 2, 2002 (Provisional program)

Start	End	Action	Title	Actor(s)
14:30	14:45	Opening	Introduction to this workshop	Steven Krauwer
14:45	15:15	Talk	Summary of the MREMS Workshop	Mark Maybury
15:15	15:40	Talk	*Challenges and Important Aspects in Planning and Performing Evaluation Studies for Multimodal Dialogue Systems*	Susanne Höllerer
15:40	16:05	Talk	*XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus*	John Bateman, Judy Delin, Renate Herschel
16:05	16:30	Talk	*Towards a roadmap for Human Language Technologies: Dutch-Flemish experience*	Diana Binnenpoorte, Catia Cucchiarini, Elisabeth D'Halleweyn, Janienke Sturm and Folkert de Vriend
16:30	17:00	**Break**		
17:00	17:30	Talk	About Roadmapping: Introduction to the plenary exercises	Steven Krauwer / Hans Uszkoreit
17:30	18:30	Exercise 1	Identifying priorities	All
18:30	19:30	Exercise 2	Putting them on a timeline	All
19:30	20:00	Discussion	Where to go from here	All & Steven Krauwer
20:00		**Closing**		

Workshop Organisers and Programme Committee

- Steven Krauwer (ELSNET / Utrecht University) (Chair)
- Hans Uszkoreit (DFKI Saarbruecken)
- Antonio Zampolli (Univ of Pisa)
- Joseph Mariani (LIMSI, Paris)
- Ulrich Heid (IMS Stuttgart)
- Khalid Choukri (ELDA Paris)
- Mark Maybury (MITRE)

Table of Contents

Author Index

Preface

Aim of the workshop:

The aim of the proposed workshop is to bring together key players in the field of resources and evaluation in order to make a first step towards the creation of a broadly supported Roadmap for Language Resources, i.e. a broadly supported view on the longer, medium and shorter term needs and priorities. This activity should be seen in the context of ELSNET's other roadmapping activities (see http://www.elsnet.org/roadmap.html), which aim at developing a technological roadmap for the whole field of Human Language Technologies.

The purpose of such roadmaps is to give the R&D community an instrument to identify opportunities for concertation of their activities and better exploitation of possible synergies between players all over the world.

Scope of this workshop:

there is no standard model for roadmaps for resources and evaluation available, we will narrow the scope of this roadmapping workshop to a specific sub-area: Multimodal Language Resources and Evaluation. This will make our discussions more focused and concrete, and it will also allow us to exploit the fact that this workshop will take place the As day after the workshop dedicated to Multimodal Resources and Evaluation of Multimodal Systems (MREMS) in general.

Recommended reading (preferably before the workshop):

- ELSNET's First Roadmap Report, edited by Ole Bernsen (http://utrecht.elsnet.org/roadmap/docs/rm-bernsen-v2.pdf),
- ELSNET's Second Roadmap Report, edited by Dorothee Ziegler-Eisele and Andreas Eisele (http://utrecht.elsnet.org/roadmap/docs/rm-eisele-v2.pdf),

Both reports can be found in the annex of the proceedings

Results of the workshop:

The results of the workshop will be published on the ELSNET website at http://www.elsnet.org

April 2002

Steven Krauwer

(ELSNET Co-ordinator)

Challenges and Important Aspects in Planning and Performing Evaluation Studies for Multimodal Dialogue Systems

Susanne Höllerer

ftw. Telecommunications Research Center Vienna
Tech Gate Vienna
Donau-City-Straße 1/ 2nd floor
1220 Vienna
Austria
hoellerer@ftw.at

Abstract

In this paper I want to discuss the problems researchers face in trying to plan and carry out an evaluation study for multimodal systems – particularly in qualifying the purpose of the testing, defining the intended user group for their application, arranging the testing setting and aligning the evaluation plan. It is my intention to show which aspects should be taken into account and which basic standards should be fulfilled. Furthermore, I provide two sections about points to consider in performing the study as well as in the analysis of the received data. Then I describe possible difficulties concerning the evaluation of (multimodal) systems and try to sketch longer term solutions. Finally, I list possible options on how to utilize the results of evaluations studies in further research.

1. Introduction

Multimodal dialog systems should be efficient, easy to handle and comprehensible for intended users – so how should the evaluation of such dialogue systems be designed and carried out in order to accomplish these goals and how can the outcome and the conclusions of the concerned studies be used for further research and development?

How can researchers and developers of dialogue systems answer the needs and preferences of the users, how can they accommodate their special interests and characteristics?

And how can problematic issues researchers and developers face today be tackled and solved? To which extent can experiences made until now help to find solutions for the future?

These questions need to be answered already in the very beginning of the whole development process – before the start of the planning and mental development of the system one has to designate the goals of the system and which functions it serves. At the same time the target group has to be defined – this can on the one hand be a small group of experts and for a special field of application or on the other hand the entire population, depending on the system or object. During the development process these facts must be taken into account in order to produce the most efficient system for the special target group. To this end, it is useful to perform an iterative mode of evaluation which means that for every important phase of development an evaluation study is provided so as to find out about the direction the development of the systems leads to and to make sure that the intended users are able to handle it. Especially as far as multimodal dialogue systems are concerned, evaluation studies are a relevant part of the development and – at the same time – a challenging task. The particular difficulty is to provide methods for logging and analysing two or more different modalities and to test each of them separately as well as combined with the other(s). This means that the developers and researchers receive much data, which require experience and reliable methods to be analysed.

Because of the innovative design and handling of those systems a careful evaluation planning has to be provided.

I wrote this paper from a social scientific point of view – as a different perspective concerning the preparation and the procedure of evaluating a system. Social and empiric science can provide information on methodological issues, questions concerning the analysis of the data, the selection of test subjects, the arrangement of the setting and the formulation of the specific tasks for the test persons.

2. Goal of the paper

The main intention of this paper is to show how important a careful planning and performance of an evaluation study – concerning especially multimodal dialogue systems – is. It should be made clear which features of the testing process are particularly relevant and which problems may appear and which challenges the evaluation of a system, possessing more than one modality for handling, may involve. Furthermore, in this paper important aspects which may appear negligible at the first sight should be mentioned, for example the range of persons who are going to use this system in the future, ergo the intended user group: what are their characteristics, their needs and how can the system serve them? For the designer and the researcher, this means also knowing exactly the functions of the system. Another aspect would be the setting in which the testing should take place: how should it be arranged and which role plays the tester?

A very important part of this paper is the one about how the results of the evaluation research in general and the experiences of each researcher can contribute to the further research done in these fields and consequently to establishing standards for the design and the performance of evaluation studies of multimodal systems.

I also like to state my point of view concerning the present as well as the longer term problems researcher might face in developing and evaluating multimodal dialogue systems, and also how they might be avoided.

3. What is the image of the intended (average) user and how does it affect the development of the evaluation of a system?

As I mentioned before, theoretically the entire population can be the target group of a certain system or object, for instance of information extracting systems like automatic telephone enquiry for train schedules. As far as IT systems are concerned, in the last years it was often assumed implicitly that the circle of intended users is a rather small one (compared to the one of objects of everyday life) and is composed of experts of fields like computer technology and science, managers or other academic job-holders in hierarchical higher positions.

But today such systems should serve everybody. It is the developers´ duty to design the system in such a way that it can also be conceived and used by non-experts. That concerns especially the presentation of the graphical user interface, which the user gets the first impression of before even having tried out how to handle the application.

So if one has in mind that the target group may be as heterogeneous as the general population there is no possibility to postulate any specific knowledge or experience concerning multimodal dialogue systems among all persons. This means that the researcher has to begin at the very start and make the use of the system as easy as possible. That is for sure a very challenging task – and an important one, because the design and the usability of the system are important factors for its acceptance among the intended users. One has to consider that persons of every age group, sex, society position and socialisation background may use this system. The sample the researcher assorts should be representative in so far that each of these parameters is taken into account. One possibility to find out about those features is to provide a user questionnaire.

One option is to search test persons of a certain age, sex and education level. The last parameter is useful to get some information about the position in society they bear. Another way would be to consider income or field of profession, or respectively the job they are working in. It is hard to find out much about the economic or social background of the test persons without violating their privacy. And one must not believe that one statistical feature gives information about a person's standard of living. So this parameter is a rather hard one to obtain. Nevertheless it should be included in the evaluation.

The aspect of age is also an important one, because one may find big differences between younger and older people concerning their competence as well as their experience with modern technological instruments and systems – a phenomenon which does not apply universally. But there may be the tendency that older persons are more sceptical and reserved if not afraid to serve as test persons for evaluations of such systems. They often argue that they need not be taken into account, because they are too old – which is of course a misbelief.

It is common to consider sex as a variable, too, because it is interesting to view possible differences between women and men in handling technological systems and to react to them in the further development of the system.

4. Recommended standards for multimodal dialogue systems

It seems to be of use to establish basic standards that need to be fulfilled in order to provide a system appropriate for a great range of users. This is particularly relevant for multimodal dialogue systems, which provide several ways of handling and therefore require extraordinary user-friendliness. The standards described in this chapter can be seen as provisional and extensible – they should serve as basic points of orientation.

These standards make clear which direction further research should take and on which aspects it should focus, but also which main issues any evaluations study should focus on.

4.1. Easy intelligibility of the functions and applicability of the system

In order to be able to use the system in an efficient way the users have to understand which actions one can perform and which goals one can accomplish with it. That is to say that the instruction manual must be clear and specific. But also the design of the user interface should give a clue to how to use the system.

4.2. Distinct visual design of the graphical user interface

The user interface, i.e. the part of the system the user sees and interacts with, should not be complex, but the various elements should be arranged clearly and distinguishably. Concerning this feature, knowledge from fields like psychology or the specific domain of advertisement could be advantageous, but the cooperation between these fields and the one of IT is not that strong yet.

4.3. Good intelligibility of the commands

The language in which the user communicates with the system is usually a set of commands – either given via speech or via GUI. And vice versa the systems gives commands or poses questions to the user – often in the form of spoken prompts. It is necessary to formulate these in a simple and intelligible way so that the user is able to catch it.

4.4. Good speech recognition

A dialogue system that provides the modality of handling via speech needs to have an excellent speech recognizer. That is a prerequisite for efficiency, which is an overall goal of such systems. This means that it should also work in noisy environment, as the system should be adaptable in awkward situations where the user cannot have regard of a clear articulation. Unfortunately – although there has been much research done in this area – it takes a long time to develop a good recognizer respectively it is hard to find a recognizer appropriate for the functions a system should fulfil

4.5. Efficiency as well as smooth performance of the actions

Multimodal dialogue systems have the special aim to work smoothly even in difficult or stressful situations, for instance if the user needs both hands for other actions. It would be very exhausting for the user to be forced to

repeat the commands or questions a several times because of the slow processing or the long upload time of the system.

4.6. Clear, intelligible output (speech-output as well as output via the GUI)

In order to provide a smooth process and a good information extraction respectively an optimum support the output the system delivers should be correct as well as intelligible.

5. Advantages of multimodal dialogue systems

The advantages listed in this chapter are supposed to supplement or – in part – condition each other. This list should – on the one hand – emphasize the differences between single- and multimodal dialogue systems and – on the other hand – show which anticipations researcher have concerning these kind of systems.

To reach the intended users as well as to make the system interesting for them, one has to emphasize its advantages in achieving a certain goal, possibly by comparing it to other kinds of systems or – in general – ways to reach this goal (for instance using a multimodal dialogue device to extract information about the surroundings instead of a simple map).

One big use of such a multimodal dialogue system is for sure the *flexibility*. The overall goal of the development of multimodal applications is for the users to interact with the system the way they like to – depending on the situation they are in. For example when driving in his car, the user cannot use his hands to operate the system – so there has to be one or more other ways to handle it in order to fulfil the claims of efficiency and usability. In this case, the modality of handling via speech input is an optimum alternative.

The optimum situation would be that every user was free to interact with the system the way the situation requires it – and to alternate the one modality with the other(s) in a spontaneous way. The system should therefore be designed to react and adapt to this user-specific behaviour. This demands – in case of a multimodal dialogue system – an excellent speech recognition as well as a synoptic user interface quick to apprehend.

Beside flexibility higher *efficiency* is another advantage of multimodal dialogue systems – provided that sufficient evaluation studies has been performed in order to find out about how a system needs to be designed to serve the users well. It´s clear that efficiency – at least in part – grows proportionally with flexibility (and the other way round), so these two aspects are connected tightly.

A third advantage which may be of great importance for "everyday users" is the *individuality and personality* a system gets when becoming multimodal, hence being able to be integrated smoothly in ones everyday life and supporting the performance of certain actions.

The great use of multimodal dialogue systems in comparison to other systems is the fact that they combine the advantages of the single modalities they include, this means that the user can profit from the advantages of handling via the GUI as well as via speech. In detail, this would be promptness as far as the modality of speech is concerned – action can be executed far more faster by speaking the commands that by typing them. The other advantage which is already known is the possibility to keep ones hands free for other actions which is, for instance, very important while driving the car. Regarding the modality of handling via the GUI the main advantage lies in the privacy of the commands the user is giving and of the actions the system is executing. While speech can be received by persons around the user, actions like typing are not audible.

Disadvantages of one modality might be eluded by using the other modality – for instance if the speech recognizer does not work properly.

6. Important items in planning and carrying out an evaluation

To evaluate a system one has to know exactly which functions it possesses and who the intended users are (cf. Nielsen 1993: 170). The evaluation study is performed to serve the purpose of finding out more about how the system should be designed in order to answer the needs and interests of the users. As I mentioned in the introduction the best way to carry out an exhaustive evaluation study is to perform several smaller "steps of evaluation". This means that – depending on the development stage of the system – the respective properties, the design and the effect on the users need to be measured.

And for each of these steps some important points must be considered. To receive sufficient and eligible data for the analysis afterwards, the evaluation study needs to be planned carefully, tasks for the test persons to perform must be formulated – which are supposed to accomplish the intentions the researchers have –, methods to log the process of evaluation need to be found as well as methods to capture the impressions and experiences of the test persons. The choice of these instruments depends on which aspects of the tests are important for the developers on the one hand, and – on the other hand – how easily the requested information can be extracted. A good way to find out which methods are appropriate for the evaluation study is to evaluate the logging methods themselves. That is also useful to assure that the methods one uses really measure what they are pretending to measure – hence if they are suitable for what the respective developer wants to find out. A good method for logging the evaluation process is to use instruments like audio recorder, video camera, mouse tracker, screen logger or eye tracker. But one must be aware that receiving too much data out of an evaluation study can be as well a problem as receiving too little.

Not only the methods and the technological equipment are to be considered – the whole setting of the testing process needs to be planned. The role of the tester who stays with the test persons must be defined – mostly he is the one who explains the aim of the testing as well as the specific tasks and who observes the test person during the performance. This raises some questions: How much information should the test person be given in order to not affect the authenticity of the situation and the (possible) impartiality of the user? Should the tester answer questions during the testing? Where should he place himself? To which extent should he adapt himself to the test person (concerning behaviour, speech, …) to provide a more informal setting or how can he prevent himself

from doing so? Are there differences in the performance of the test persons depending on the sex, the age or the credibility of the tester?

As far as the test persons are concerned, should they be given some time to get to know the system better (some minutes without logging or even observing) or should they be tested from the very beginning?

And how should the testing setting look like to provide as much authenticity as possible?

All these questions can become problematic when too little time and know-how is spent on the preparations of the evaluation studies – the difficult issues are explained in detail in chapter 7.

7. Analysis methods

It is important to find appropriate analysis methods as well, for instance annotating schemes to analyse spoken language and synchronize it with actions like mouse movements or clicks. That is a good and rather objective possibility of spotting the problems the users had performing the tasks, but also the points where the test persons apparently used the system in an efficient way, for example – concerning multimodal dialogue systems – combined speech and handling via the GUI. Especially for large numbers of test persons and hence a lot of data such standardized analysing methods are useful. However, the range of good and reliable annotating schemes is not that great. The few that are made use of in empirical studies fall short of easy applicability and efficient programming. Much remains to be done in this field of research. Also the methods themselves need to be tested to find out if they work the way the researcher wants them to. If the methods fail, the whole study needs to be repeated.

However, also the subjective impressions of the test persons are important for the analysis, so one should not surrender a questionnaire, an individual interview or informal talk with the test persons after the testing. These data need to be analysed either quantitatively – in the case of a questionnaire – and presented in statistics or analysed in a qualitative way, that is to collect the test persons' impressions and statements and to detect positive or negative tendencies.

But not in every case all the errors of a system can be detected: one cannot be sure that all the problems could actually be recovered – "one troubling aspect of testing is the uncertainty that remains even after exhaustive testing by multiple methods". [Shneiderman 1998: 125]

8. Problems to be solved concerning the process of developing and testing a new system

There is a range of problems researchers of multimodal dialogue systems have to face during the process of developing and optimizing the system. In some ways, preparing the evaluation study for multimodal dialogue systems does not differ from preparing one for "singlemodal" systems. Just a few aspects are more challenging as far as multimodal systems are concerned.

8.1. Defining the user group and test subjects

First of all, it is difficult to find out about the intended user group: how should the researchers know which persons the system will be used by? And how can they be sure that the users they design the system for are really the ones who will use the system in the end? A step towards finding a solution to this problem would be to carry out a survey among the supposed target group or among the whole population to get a clue about who is interested in the product and may benefit from it.

The range of test persons should be representative for the group of intended users, that is to say that the test subjects should represent the properties of the target group. If the system was designed primarily for elder persons, it is recommended to choose such persons for the evaluation study. In this regard the question must be raised where one should find appropriate test subjects. There are several possibilities:

One may look for persons in public institutions or buildings like schools, universities or on the street. An alternative would be the search for people by an advertisement. Or one may get access to a range of test persons by buying (or exchanging) subjects databases.

8.2. The discrepancy between researcher and user

Another difficulty in the process of developing a (multimodal dialogue) system is the discrepancy between researcher/developer and "normal" user or test subject. The researcher who designs the system is an expert in this field, he/she possesses knowledge and experiences concerning this specific system and knows how to handle it – so one can assume that he/she is the person appropriate for testing the system. That is true – to some extent. The persons, who understand the functions and operations of the system best, may also know how to measure and optimize them. The problem, which may occur, is that the researcher knows the system to well. This means that he/she is not able to put him/herself in the situation of the non-expert user and, therefore, blind out all his/her knowledge. One may argue that just because of these problems evaluation studies are carried out. That is correct. But it is not enough to perform one or more evaluation studies, it has to be guaranteed that the study is performed in a correct way, this means to really find out about the user group and its needs and expectations. The researchers are – in some way – preoccupied. So they do not seem to suit for planning such a study. A possibility to avoid this problem would be to separate the role of the researcher and the one of the evaluation designer strictly. But here another difficulty appears: how can the evaluation designer know enough about the system to understand its functions and features and at the same time know not too much about it in order to stay as objective as possible?

8.3. The evaluation setting

In order to get valid testing results, not only the tasks to fulfill need to be chosen carefully, also the setting where the testing should take place has to be planned regardfully.

The easiest possibility is to perform the tests within an isolated laboratory or at least in the rooms of the company which developed the system. This would mean that the testing situation could be controlled rather easily and that no unexpected disturbances would happen. These apparent advantages entail one negative aspect. Choosing such a testing setting would mean that the authenticity of the situation would be in peril. Especially as far as

applications are concerned which are not designed to be used at home or in a quiet and private place, testing within the circumstances mentioned above would not represent the conditions which the user has to face when using the application in reality. The researcher cannot foresee all the different situations in which the system may be applied but he knows the intended user group and the functions of the application and therefore can assume how it is going to be used.

Portable devices for example are supposed to be applied on the way, for instance on the streets, in public buildings and institutions, while walking or traveling by car, at different events or in likewise noisy environment. The noise must not be underestimated – as well as other factors, for instance when information is required as quick as possible (train departure times for example). A system, which works perfectly within the laboratory setting, might turn out to fail when being used in real surroundings. How should these settings be imitated in the laboratory to gain valid results?

As a matter of course, one must in this case consider the development phase of the system. If there is not an application to be carried around yet it can hardly be tested like if there was. An iterative kind of evaluation study requires several different testing settings.

8.4. Methods

Finding appropriate methods for logging the testing process might be a problem as well. The choice depends on which modalities the system has, as there are several options for each of them to be logged and measured. Most multimodal dialogue systems offer at least the two following modalities: the speech-modality and the handling via the GUI.

In order to log spoken user-output, one could for instance use a simple recorder with a microphone or a camera which could also tape visual impressions like the gesture and the face of the test subject as well as the monitor of the computer or the display of the application device (if it is not too small) – depending on which kind of system is tested. At the same time, the output of the system should also be taped for to liaise the both kinds of output in order to get information about the quality of the speech recognition and the smoothness of the whole process.

It is not as simple to find methods – beside the camera – to trace the operations on the monitor or the display, ergo the handling via the GUI. There exist some software tools like screen logger or key tracker which log the mouse movements or clicks as well as the input via the keyboard or the selection via the menu. Unfortunately, the existing software is either very expensive or only available for companies of specific fields.

In addition, there are other tools to log the process in order to gain more information about the handling of the system – for instance a so-called eye-tracker that logs the eye movements of the user. Through its analysis one may find out about which elements of the GUI are bold and how easy or difficult it is for the user to understand how to operate the system.

The challenging aspect concerning multimodal systems is to connect the methods used for logging the handling of different modalities. One kind of information needs to be related with another. The speech signals must be synchronized with the manual actions, for instance. This intention requires another software or program like an annotation scheme.

8.5. Possible solutions and recommendations for the future

As I said before, it is necessary to involve several persons in the developing and the testing process of the system, as more perspectives are required for an effective evaluation study. Concretely, this means that persons from several fields of research should work together, the tasks should be distributed and the roles the persons occupy within this process should be defined well. The researcher, the developer, the designer, the market research institute, the tester, several university institutes like psychology, sociology and computer science – all of these persons and institutions have competences in their specific fields and can contribute to producing a good working system. Through exchanging experience and know-how, as many difficulties as possible might be avoided.

In my view, this strategy will play an important role in the future, for aspects like user friendliness and acceptance of the system by the users are more and more coming into prominence. It is not any longer the group of IT experts and business people only who need applications of new technologies, but “average persons” from every part of the society.

Nowadays the number of those companies increases which specialize on evaluation studies and tests on usability – a fact that indicates the prominence of these aspects.

While IT companies spent most time on producing new systems and optimising new technologies, the aspect of user friendliness was rather neglected. The big chance to catch up on these experiences is the cooperation with persons of other fields or companies; to get support at finding the right test subjects, equipment and methods.

9. Evaluation outcomes as resources for further research

First of all, the outcome of the evaluation studies serves the improvement and the development of the evaluated system. But the received data are not useless after completing the evaluation process. The lessons one draws out of this testing can be used for other – similar – studies. On the one hand developers get to know the logging and analysing methods better, on the other hand they learn more about this kind of systems in general and how the intended users manage them – this knowledge can be made use of in further research.

To mention the economical aspect, the results of an evaluation study can of course also be used in cooperation with other technological enterprises or research centers with commercial as well as scientific interests; they can be exchanged or sold.

This procedure does not need to be restricted to similar, ergo technological, fields of research – instead the knowledge can also be connected to different fields like psychological or sociological research or the particular field of advertisement, where methods of usability and analyses of effects on consumers and users have long tradition. Knowledge from these disciplines can be used for evaluation studies and – vice versa – the results of these kinds of studies can be made use of in other fields.

10. Conclusions

The challenges in designing evaluation studies for multimodal dialogue systems are plain to perceive: in contrast to dialogue systems using one modality, evaluating multimodal systems demands more than one perspective of testing and hence just as many methods of logging. To use the received data for the improvement of the system and for further research it has to be processed with the support of suitable analysis methods – the particular challenge at this is to find an appropriate method for each modality and each kind of data.

Another aspect is the definition of the purpose as well as the intended users, and as a main task the designing of the user interface and the systems functions in order to meet the interests and needs of the target group.

The field of research of multimodal dialogue systems and applications is relatively new and few standards concerning the design of the user interface or the ways of testing and analysing have been established. But today usability studies are attached more importance than ever – for every kind of system or object, not only in fields of technology. There are – on the contrary – branches that deal frequently with aspects of usability and already gained precious information, for instance (cognitive) psychology. This knowledge can be useful for enterprises or persons who develop such multimodal systems. In my opinion, the cooperation with other companies or even other fields of research and hence the exchange of experiences and know-how is one big chance to improve usability testing, even for very specific applications.

Acknowledgements:
This work was supported within the Austrian competence center program ***K****plu*s, and by the companies Kapsch, Mobilkom Austria and Siemens.

11. References

Nielsen, Jakob (1993): Usability Engineering. Academic Press. San Diego.

Shneiderman, Ben (1998): Designing the User Interface. Strategies for Effective Human-Computer Interaction. Addison-Wesley. Reading, Massachusetts.

XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus

John Bateman*, Judy Delin†, Renate Henschel‡

*University of Bremen, Bremen, Germany
bateman@uni-bremen.de

†University of Stirling, Stirling, Scotland
and Enterprise Information Design Unit, Newport Pagnell, Bucks, England
j.l.delin@stir.ac.uk *and* judy.delin@enterpriseidu.com

‡University of Stirling, Stirling, Scotland
rhenschel@uni-bremen.de

Abstract

Current views of multimodal language resources have not yet sufficiently captured the complex interrelationships within page-based information delivery. This is critical for development of multimodal corpora and language resources suitable for large-scale empirical investigation. Serious attempts to interrogate the nature of multimodal meaning-making in professionally-produced documents, both paper and electronic, require a clear understanding of the organisation of the layers into which meaning is organised. In this paper, we present the first multi-layered XML annotation scheme that meets these requirements, developed using a combination of expertise from computational linguists and designers from various sectors of the publishing industry.

1. Introduction

With current developments and goals involving multimodal documents in the widest sense—i.e., including highly interactive artifacts capable of responding to, and producing information in, input/output modes ranging across verbal, gesture, touch and so on, animated/video content, traditional texts, graphics, and so on—it is perhaps tempting to believe that the organization of 'simpler', more traditional document forms, such as two-dimensional presentations involving textual, graphical and diagrammatic information, has been 'solved'. Attention is then drawn away from the complexities of these document types, such as they are, and are to be picked up as a by-product of dealings with more complex artifacts. In our ongoing work on two-dimensional, non-animated information presentations—e.g., books, information leaflets, traditional websites, newspapers (in both print and online forms), and so on—we have found a wealth of complexity that raises serious doubts about such an approach. One aspect of the problem, and the challenge, can be seen in the large gap that exists between previous corpus encoding initiatives (e.g., TEI and the derived CES) which are text based and more recent proposals for capturing mixed media/mode presentations: Somewhere between these two extremes, much of the highly flexible and meaningful resources of two-dimensional information presentation traditionally and non-technically subsumed under 'layout' and graphic design go missing.

As a consequence of this, we have found it necessary to develop a new annotation scheme for describing the informational relationships employed in the area. Two-dimensional information presentation—whether on the page, screen, or whatever—still represents the overwhelming majority of users' contact with information, and so a revealing and empirically based understanding of the meaning-making resources of this area remains of crucial importance. Previous attempts to provide annotation schemes for setting up corpora for documents of this kind have not succeeded in covering very much of the range of phenomena encountered in natural documents however (Corio and Lapalme, 1998; Bouayad-Agha, 1999; Bouayad-Agha, 2000). In this paper, we describe the goals of our own annotation work, set out the basic levels of annotation we believe are required, describe the technical approach taken, and indicate what we see as the next immediate stages, problems and challenges of follow-up development.

2. Goals

We take the view that language, layout, image, and typography are all purposive forms of communication. Accordingly, in our research project GeM ("Genre and Multimodality", http://www.purl.org/net/gem), we aim to describe and analyse all these elements within a common framework, thereby providing a more complete understanding of meaning-making in visual artefacts. By analysing resources across visual and verbal modes, we can see the purpose of each in contributing to the message and structure of the communicative artefact as a whole.

One particular goal of the research is to formalise and model the role of *genre* in layout and typographical decisions. Through the analysis of sample types of multimodal document, the project aims to develop a theory of visual and textual page layout in electronic and paper documents that includes adequate attention to local and expert knowledge in information design. The model is being implemented in the form of a computer program that allows exploration of both existing and potential layout genres, generating alternative and novel layouts for evaluation by design profes-

sionals.

Our use of the term genre here is similar to Biber's (1989, pp5–6), who in his study of linguistic variation states that 'text categorizations readily distinguished by mature speakers of a language; for example—novels, newspaper articles, editorials, academic articles, public speeches, radio broadcasts, and everyday conversations—categories defined primarily on the basis of external format'. We adhere, too, to Biber's view that these categories of text also reflect distinctions in the author's purpose: the documents look different, and contain different language forms, because they are intended to do different things.

Although there are many attempts to categorise the kinds of language that occur in different genres of texts in linguistics, there are few attempts to extend genre analysis into other aspects of visual meaning: Twyman (1982) and Bernhardt (1985), for example, provide preliminary schemes for categorising documents according to the interrelationships between images and text, while Kress and Van Leeuwen (2001) have now also explicitly begun to relate multimodality and genre. Waller (1987), however, is the only attempt extant, to our knowledge, that has attempted to describe the role of language, document content, practical production context and visual appearance in the formation of document genre within the same framework. Our work draws upon and extends Waller's in several ways, as we shall make clear below.

For this, or any project addressing the communicative strategies involved in two-dimensional visual artefacts, the provision of suitable corpus materials is fundamental. Furthermore, since such materials are not currently available, the development of such a corpus has been adopted as an additional explicit goal of the GeM project. The purpose of the corpus development within GeM is to investigate systematic connections between a rich characterisation of the context of use of multimodal documents and their linguistic, graphical, and layout realisations. Within the GeM project itself, four broad document genres have been selected for initial treatment: traditional paper-based newspapers, online web-based newspaper sites, instructional documents, and wildlife books; in each area we have secured a collection of documents and have established contact with designers either expert in these respective fields or, in several cases, actually responsible for the documents gathered. We focus here on the annotation scheme that we have found necessary for structuring the corpus developed.

3. Basic levels of annotation

Waller (1987, pp178ff) represents the constraints on the typographer in producing a graphical document as emerging from three sources:

- Topic structure: 'typographic effects whose purpose is to display information about the author's argument—the purpose of the discourse';
- Artefact structure: 'those features of a typographic display that result from the physical nature of the document or display and its production technology';
- Access structure: 'those features that serve to make the document usable by readers and the status of its components clear'.

Waller did not produce detailed text analyses based on his model but, grounded as it is in the very practical concerns of document design, his view that document appearance results from satisfying goals at different levels is persuasive. We have particularly taken the force of his point that the physical nature of the document and its method of production play a major role in its appearance. In this way, the 'ideal' layout of information on a page may never occur: it must be 'folded in' to the structures afforded by the artefact, and labelled and arranged according to the structures required for access. Document design is therefore never 'free', in the sense that it is never motivated solely by the dictates of the subject matter. We therefore have required a place for these kinds of constraints in our annotation.

In our revision of Waller's model, we suggest that there is an advantage to be gained in uncollapsing his 'topic structure' into a separation between content and rhetorical presentation. We view content to be the 'raw' data out of which documents are constructed. What Waller describes as 'the author's argument' is not solely or completely dictated by content: many rhetorical presentations are compatible with the same content. In terms more familiar from natural language generation, we separate out the 'what-to-say' from rhetorically structured text plans for expressing that content. Secondly, we take what Waller terms 'artefact structure' to be not a structure in the sense of some set of ideas that are to be incorporated in the document, but rather as a constraint on the combination of all the other elements into a finished form.

The levels we propose as minimally necessary for revealing accounts of the operation of the kinds of visual artifacts being gathered in our corpus are, then, as follows:

- Content structure: the structure of the information to be communicated;
- Rhetorical structure: the rhetorical relationships between content elements; how the content is 'argued';
- Layout structure: the nature, appearance and position of communicative elements on the page;
- Navigation structure: the ways in which the intended mode(s) of consumption of the document is/are supported; and
- Linguistic structure: the structure of the language used to realise the layout elements.

We suggest that document genre is constituted both in terms of levels of description, and in terms of the constraints that operate on the information at each level in the generation of a document. Document design, then, arises out of the necessity to satisfy communicative goals at the five levels presented above, while also addressing a number of potentially competing and/or overlapping constraints:

- Canvas constraints: Constraints arising out of the physical nature of the object being produced: paper or

screen size; fold geometry such as for a leaflet; number of pages available for a particular topic, for example;

- Production constraints: Constraints arising out of the production technology: limit on page numbers, colours, size of included graphics, availability of photographs; for example, and constraints arising from the micro-and macro-economy of time or materials: e.g. deadlines; expense of using colour; necessity of incorporating advertising;

- Consumption constraints: Constraints arising out of the time, place, and manner of acquiring and consuming the document, such as method of selection at purchase point, or web browser sophistication and the changes it will make on downloading; also constraints arising out of the degree to which the document must be easy to read, understand, or otherwise use; fitness in relation to task (read straight through? Quick reference?); assumptions of expertise of reader, for example.

Following Waller (1987), then, we claim that not only is it possible to find systematic correspondences between these layers, but also that those correspondences themselves will depend on specifiable aspects of their context of use. In particular, they will depend on 'canvas constraints' set by the nature of the realizational medium (paper, screen-based browser, palmtop, screen resolution) and 'production constraints' imposed by available technology and design choices (allowable cost, number of pages, available printing or rendering techniques, etc.). A model of multimodal genre must begin by expressing adequately the above five levels of description as well as finding the most appropriate way of satisfying the three sets of constraints.

Our provision of a corpus of multimodal documents serves as the empirical basis for more thorough investigations of this claim. So far our work has identified widespread mismatches between rhetorical purposes and layout structures even among professionally produced documents; this offers a useful basis for constructive critique. We see the collection of extensive corpora of multimodal documents of this kind, annotated according to the levels of description that we have here briefly motivated, as an essential research and direction for the next five years.

4. Technical implementation

As we have seen, the two communication modes of visual and verbal information presentation are the main perspectives to be captured in the GeM annotation scheme. The scheme accordingly identifies textual elements (verbal mode) and layout elements (visual mode) in a multi-layered annotation, and specifies how these elements are grouped into hierarchical structures (primarily: the rhetorical structure for textual elements, the layout structure formed by the layout elements, and a page model formed by an 'area model': see below). The alignment between these intersecting hierarchies is achieved by specification of the 'GeM base'—a list of the basic units out of which the document is constructed. In accordance with the goal of the GeM project, the granularity of the linguistic basic units employed in the annotation is approximately the sentence level—this does not preclude providing correspondences with other levels of granularity that might be required for other purposes of course.

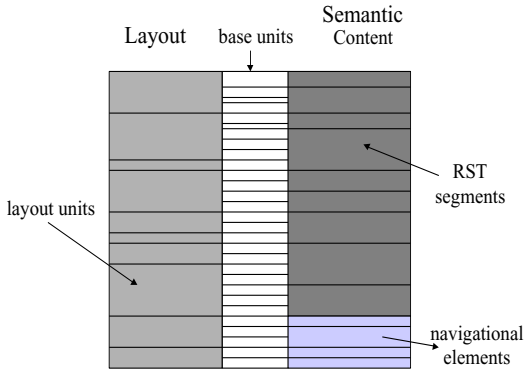

Figure 1: The distribution of base elements to layout, rhetorical and navigational elements

Each layer in the GeM model is represented formally as a structured XML specification, whose precise informational content and form is in turn defined by an appropriate Document Type Description (DTD).[1] The markup for one document then consists generally of the following four inter-related layers:

Name	content
GeM base	base units
RST base	rhetorical structure
Layout base	layout properties and structure
Navigation base	navigation elements and structure

All information apart from that of the base level is expressed in terms of pointers to the relevant units of the base level. This stand-off approach to annotation readily supports the necessary range of non-isomorphic, overlapping hierarchical structures commonly found even in the simplest documents. The relationships of the differing annotation levels to the base level units is depicted graphically in Figure 1. This shows that base units (the central column) provide the basic vocabulary for all other kinds of units and can, further, be cross-classified.

This annotation scheme is being developed further in response to the needs of concrete annotation tasks. Its current state is described in the technical manual available on the GeM website (Henschel, 2002). We describe it further here only in sufficient detail to give an impression of the kinds of annotation information and work involved.

4.1. Basic constituents

The purpose of the base level annotation is to identify the minimal elements which can serve as the common denominator for textual elements as well as for layout elements. Where speech-oriented corpora use the time line as basic reference method, and syntactically oriented corpora use the sequence of characters or words, the GeM annota-

[1] For the DTDs themselves, as well as further information and examples, see the GeM corpus webpages.

tion operates at a less delicate level and uses bigger chunks (mostly sentences and graphical page elements) as the bases of the markup. Everything which can be seen on each page of the document has to be included. How the material on each page is broken up into basic units is given by the following list, each is marked as a base unit:, orthographic sentences, sentence fragments initiating a list, headings, titles, headlines, photos, drawings, diagrams, figures (without caption), captions of photos, drawings, diagrams, tables, text in photos, drawings, diagrams, icons, tables cells, list headers, list items, list labels (itemizers), items in a menu, page numbers, footnotes (without footnote label), footnote labels, running heads, emphasized text, horizontal or vertical lines which function as delimiters between columns or rows, lines, arrows, and polylines which connect other base units.

Everything on a page should belong to one base unit. The base annotation has a flat structure, i.e. it consists of a list of base units.[2] Generally any text portion which is differentiated from its environment by its layout (e.g. typographically, background, border) should be marked as a base unit. The list of base units needs to comprise everything which can be seen on the page/pages of the document. The tag used to mark base units is the <**unit**>. Each base unit has the attribute **id**, which carries an identifying symbol. If the base unit consists of text, the start and end of this text is marked by the <**unit**> tag. Illustrations, however, are not copied into the GeM base. Thus, base units which represent an illustration or another graphical page element are empty XML-elements but can optionally be equipped with an **scr** and/or an **alt** attribute to show, indicate or access the source of an illustration.

4.2. Layout base

The layout base consists of three main parts: (a) layout segmentation – identification of the minimal layout units, (b) realization information – typographical and other layout properties of the basic layout units, and (c) the layout structure information – the grouping of the layout units into more complex layout entities. We explain these three components in more detail below.

In typography, the minimal layout element (in text) is the glyph. In GeM, however, we are primarily concerned with typographical and formatting effects at a more global level for a page; therefore we do not go into such detail, instead considering the paragraph as minimal layout element. That means, a sequence of sentences with the same typographical characteristica which makes up one paragraph is marked as one layout unit. In addition to that we mark all graphically realized elements from the GeM base as layout units. Also highlighted text pieces in sentences, or text pieces within illustrations are marked as layout units. Hence the same list which has been given for the markup of the base units applies here, but with paragraphs instead of orthographic sentences. The tag for a layout unit is <**layout-unit**>. Each layout-unit has the attribute **id**, which carries an identifying symbol, and the attribute **xref** which points to the base units which belong to this layout unit.

The second part of the layout base is the realization. Each layout unit specified in the layout segmentation has a visual realization. The most apparent difference is which mode has been used – the verbal or the visual mode. Following this distinction, the layout base differentiates between two kinds of elements: textual elements and graphical elements marked with the tags <**text**> and <**graphics**> respectively. These two elements have a differing sets of attributes describing their layout properties. The attributes are generally consistent with the layout attributes defined for XSL formatting object and CSS layout models.

Some of the layout units identified in the segmentation part of the layout base can be grouped into larger layout chunks. For instance, the heading and its belonging text form together a larger layout unit, or the cells of a table form the larger layout unit "table". The criterion for grouping layout elements into chunks is that the chunk should consist of elements of the same visual realization (font-family, font-size, ...), or the chunk is differentiated as a whole from its environment *visually* (e.g. by background colour or a surrounding box). In Reichenberger et al. (1995), the authors propose identifying layout chunks by applying a decreasing resolution to the document. The grouping into chunks usually can be applied in several steps, thus forming larger and larger layout chunks out of the basic layout units up to the entire document. Note that one chunk can consist of layout elements of different realizations (text and graphics). The third part of the layout base then serves to represent this hierarchical layout structure. Generally we assume that the layout structure of a document is tree-like with the entire document being the root. Each layout chunk is a node in the tree, and the basic layout units, which have been identified in the segmentation part of the layout base, are the terminal nodes of that tree.

Area model. Each page usually partitions its space into sub-areas. For instance, a page is often designed in three rows – the area for the running head (row-1), the area for the page body (row-2), and the area for the page number (row-3) – which are arranged vertically. The page body space can itself consist of two columns arranged horizontally. These rows/columns need not to be of equal size. For the present, we restrict ourselves to rectangular areas and sub-areas, and allow recursive area subdivision. The partitioning of the space of the entire document is defined in the **area-root**, which structures the document (page) into rectangular subareas in a table-like fashion.[3]

The tag to represent the area root is <**area-root**> The tag to represent the division of a sub-area into smaller rectangles is <**sub-area**>, this shares the attributes of the root but adds a **location** attribute so that subareas are positioned relative to their parent. Locations are indicated with respect to a logical grid defining rows and columns. If, for example, we were considering a page made up of a running

[2] In certain cases, we diverge from the flat structure of the base file. See the technical documentation for further details.

[3] Note that the area-root need not to be a page; if the document to be annotated is a book or brochure, then it can also be the entire book or brochure.

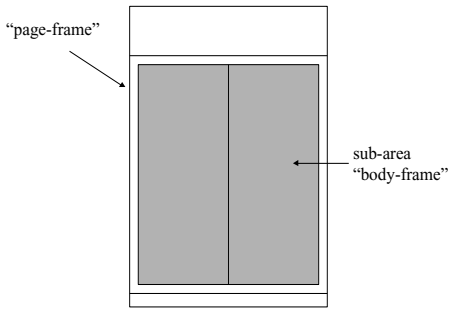

Figure 2: Visualized area model

head, a page body, and a footer for the page number, and in which the page body itself is divided into two columns, then the following annotation would define a corresponding area model. Here, the example's area model consists of a specification of the area-root (called "page-frame"), and the specification of one particular sub-area located in row-2 (called "body-frame"):

```
<area-root id="page-frame" cols="1" rows="3"
      hspacing="100" vspacing="10 85 5"
      height="16cm" width="14cm">
  <sub-area id="body-frame" location="row-2"
      cols="2" rows="1" hspacing="50 50"
      vspacing="100"/>
</area-root>
```

The attribute `vspacing=''10 85 5''` means that the running head takes 10% of the entire page height, the page body 85% and the page number 5%. The page body consisting of two columns is indicated by the hspacing attribute value "50 50", i.e., that both columns are equal in width and take half of the parent unit's width.[4] This area model is visualized in Figure 2.

The area model then provides logical names for the precise positioning of the layout units identified in the layout structure proper.

4.3. RST base

The RST base presents the rhetorical structure of the document. The rhetorical structure is annotated following the Rhetorical Structure Theory (RST) of Mann and Thompson (1988). In RST, a **span** is a continuous text fragment consisting either of a nucleus and one or more satellites (mononuclear relation), or of a number of nuclei which stand in a multinuclear relation (joint, sequence, ...) Some characteristics of RST vary between different research traditions, especially the granularity of the segmentation, the assumed set of rhetorical relations and the branching style of the rhetorical structure tree. We have also needed to make some extensions for the particularities of dealing with mixed verbal and visual information; clearly, when one wants to apply RST to modern, often multimodal, documents, new issues arise. Previous generalizations of RST to multimodal documents have either added new relations to model the relations between graphics and text (Schriver, 1996; Barthes, 1977) or parameterize the existing relation set by a mode parameter (André, 1995). We favour the second approach. However, there are other problems when generalizing RST to multimodal documents, which have not been addressed previously:

- The prominence of graphics in multimodal documents makes it often difficult to decide upon nuclearity in multimodal relations.
- The linear order of the constituents of the document is lost.
- The minimal unit for RST segmentation cannot be restricted to a clause or clause-like phrase.

We address these concerns briefly in turn.

Nuclearity in multimodal relations. Although graphical illustrations are often used to *rephrase* a text passage, it is often difficult to decide which of the two segments – the illustration or the text passage – is in fact nulear and which is the satellite. This seems to be a particular problem of graphics-text relations. To model this problem, we use the multinuclear **restatement** relation. A similar relation can also be found in Barthes under the name **redundant**.

Linear order. Conventional RST builds on the sequentiality of text segments. Relations are only possible (with some minor exceptions) between subsequent segments/spans (sequentiality assumption). With multimodal documents, the mutual spatial relations between the segments changes (from relations in a string-like object to relations in a graph). Segments can have not only a left and a right, but also an upper and a lower neighbour segment. In general one can imagine neighbouring segments in any direction, not only the four which presuppose a rectangular-based page layout. In addition to this, there can be more than one neighbour in each direction. The simpliest solution to apply RST (with its sequentiality assumption) to such a document would be to introduce a reading order on the segments of the document, which is then used as the sequence behind the RST structure. However, this can easily fail to reflect the actual reading behavior. A better, more straightforward generalization of the sequentiality assumption, which we will adopt here, is to restrict RST relations to pairs (sets) of document parts (segments/spans) which are adjacent in any direction. But again, in real documents, one can sometimes find a layout where the rhetorical structure obviously is in conflict with this adjacency condition. Our hypothesis here is that this is generally possible, but that in such a case an explicit navigational element is required so as to indicate the intimate relation between two separated layout units.

Clause as segment. The clause usually serves as minimal unit in RST. There are also approaches, which allow prepositional phrases to be a segment on their own. This is straightforward because both approaches assume something which denotes an action, an event or a state – also called eventualities – as the basic unit. However, if we move to modern documents, particularly multimodal documents, it is questionable whether the clause/PP basis should be kept. Typical examples in multimodal documents are:

[4] For the time being, we ignore space for margins, at least as long as they do not contain footnotes or other text.

- a diagram picturing a certain object and a text label which identifies (puts a name to) this object
- a list with an initiating sentence fragment, as in:

In the box are: ◇ three cordless handsets ◇ the base unit ◇ a mains power lead with adapter ◇ a telephone line cable ◇ two charger pods

- an attribute-value table, as in:

Juvenile	Grey-brown, flecked becoming whiter, adult plumage after three years.
Nest	Mound of seaweed on bare rocky ledge.
Voice	Harsh honks and grating calls at colony.

The cited examples are all expressions of states, or of static relationships between two objects or between an object and a property such as: identification, location, possession, and predication relations. In a traditional linear text, such relations would have been expressed as *is*- and/or *has*-clauses. Each such clause would constitute **one** basic RST segment. In our examples above, however, the two constituents of such a static relation clause are broken out and printed as separate layout units—in the first example, they are even given in differing modes. It is their mutual arrangement on the page plus possible extra graphical devices that expresses the relation between them. This raises the question as to what counts as a minimal unit for an RST analysis in such documents. We solve this issue by introducing a new component for annotation distinct from RST: we analyse the object-object/property relations, if they are clearly separate layout units, according to a small set of relations based on Halliday (1985), which we term 'intraclausal-relations'.

The tag used to mark the basic RST units is <**segment**>. In order to find out which base units form segments, one has to filter out those base units which are in the document for navigational reasons only. These are, for example, page numbers, running heads, footnote labels, document deictic expressions. We also consider headings as navigational elements, and do not include them in the RST analysis. In addition to these segments, we compose other complex segments consisting of more than one base unit for the cases where a intraclausal-relation is expressed on the page by two (or more) separate layout units. Typical examples are diagram + label, table cell$_{i,1}$ + table cell$_{i,2}$ in a two-column table, list initiating sentence fragment + list items. And, finally, sentences disrupted into two base units by page/column breaks only form one segment in the RST base.

The GeM XML annotation for RST aims to overcome some drawbacks found in existing RST annotation approaches. The two standards common in the RST community are those provided by the annotation tools of Daniel Marcu and Mick O'Donnell (see, e.g., www.sil.org/~mannb/rst/toolnote.htm). In both these tools, the annotated output is primarily seen as the program-internal representation of RST structures to be visualized as graphical trees with the help of the tool, but not as output to be used for further XML processing; we describe the pros and cons of the alternatives more in the technical documentation.

4.4. Navigation base

Navigation in a document is performed with the help of pointers, text pieces which tell the reader where the current text, or 'document thread', is continued or which point to an alternative continuation or continuations. The addresses used by such pointers are either names of RST spans or names of layout chunks. For long-distance navigation, typical nodes in the RST structure and in the layout structure have been established for use in pointers; in particular, chapter/section headings are names for RST spans and page numbers are names for page-sized layout-chunks, which tend to be used for navigation. However, there can also be other name-carrying layout-chunks or RST spans such as, for example, figures, tables, enumerated formulas, and so on. The navigation base of a document lists all these "names" which have been defined in this document to be actually or potentially used in pointers. We call the names of RST spans **entries** because they are usually placed immediately before the text of this span. We call the name of a layout-chunk an **index**.

The tag for an entry definition is <**entry**>. We allow entries simultaneously to be segments. We annotate the definition of an index at the page where it is defined, and refer with xref to the base unit which serves as the identifier.

Beside the list of entries and indices, which just defines addresses, the most important part of the navigation base consists of all pointers occuring in the document. The surface realization of pointers are "document deictic expressions", a term coined by Paraboni and van Deemter (2002). Document deictic expressions occur either within sentences or as separate layout units. We have marked the first type as embedded base units and the second as main level base units in the GeM base. In the navigation base, we specify the semantic meaning of such a document deictic expression as **pointer**. We distinguish pointers which operate on the layout structure, and pointers which operate on the RST structure. A pointer (or link) operating on the RST structure points from the current segment (which entails the document deictic expression) to an RST span – the goal RST span – which is layouted at a different place and is not adjacent. A pointer operating on the layout structure points from the layout chunk (which entails the document deictic expression) to another layout chunk which is not adjacent. Another distinction is the pointer type, which indicates different pointing situations. A **continuation** pointer is used in the situation where the layout of an article is broken into two non-adjacent parts. The second part is often printed several pages later than the first part. Continuation pointers are typically layout-operating pointers. **Branching** pointers are used in the situation where a certain piece of information is with respect to its content appropriate at two (or

more) places in the same document. The designer has decided to put it at one of the possible places. In order to indicate the other possible place, a pointer is given at the other location. A third type of pointer is the **expansion pointer**. It is used when more information is available, but not central to the writer's goal. An expansion pointer points to this extra information. Coming along a branching or an expansion pointer, the reader has the choice between two alternatives to continue reading the document. With a continuation pointer there is only the choice between reading continuation or stopping.

4.5. Uses of the corpus

The main results found so far in use of the corpus have been local, in that we are uncovering the rather wide variation that exists between selected layout structures on the one hand and rhetorical organization on the other *within single documents*. In surprisingly many cases, this variation goes beyond what might be considered 'good' design: in fact, we would argue that such designs are flawed and would be improved by a more explicit attention to the rhetorical force communicated by particular layout decisions. This represents the use of the corpus for document critique and improvement (cf. Delin and Bateman (2002)); here further corpus collection is nevertheless essential in order to map further the limits of acceptable functional variation.

We are also exploring the formulation of constraints over collections of corpus entries—e.g., over the pages of a book, or over collections of books in a series, etc.—by means of further annotation levels in which values from the primary annotation levels are partially specified. These need to be hierarchically related. It is at these 'meta' levels that the role of Waller's production and canvas constraints become particularly clear. We are employing this information as an important source of input in a prototype automatic document generation system capable of producing the kinds of variation and layout forms seen in our corpus, thus extending the early generation work in this spirit presented in Bateman et al. (2001).

Finally, we are still searching for more effective means of interogating the corpus maintained in the GeM style. Queries expressed in the XML Xpath language allow simple retrieval of information maintained in the corpus, but are cumbersome for more complex queries. Whether further developments such as XQL or XQuery will bring benefits is not yet clear. Somewhat disappointing was the unsuitability of the previous generation of linguistic-oriented corpus tools, which, despite considerable investment, seem to have been outstripped by the very rapid developments seen in the mainstream XML community. Most of our current work is done directly with XMLSpy and XLST tools such as Xalan. We have found the non-linearity and the non-consecutive nature of the units grouped within our annotation scheme as presenting a major problem for annotation models that have been developed in the speech processing tradition where contiguity of units is the expected case.

5. Follow-up goals, challenges and requirements

We expect that the details of annotation will be refined further as we approach a wider range of documents. It is now a major challenge to produce workable annotation schemes and corresponding corpus collections that include the kind of information we have argued to be necessary in this paper. This information represents a crucial bridging between technicalities of document production and the real issues of design faced in the publishing industry. Corpora built in this way will face two-ways: both to further linguistic and computational plinguistic research and development and to practical issues of design and evaluation. We believe that this needs a firm place in any roadmap now envisaged for language resource construction.

With this in mind we are also exploring a second round of corpus collection and annotation; it is our conviction that only a thorough corpus-oriented study of documents will allow further motivated theoretical and practical statements to be made about the meaning resources that such documents offer. If language resources are to be constructed that include documents of the kind targetted within GeM, then information such as that captured in the GeM annotation scheme will be crucial.

Here there are several issues that require concerted effort. Theoretically, the acceptance of the value and role of rhetorical analyses as giving a fine-grained description of communicative intentions is not uncontroversial. There are attempts in progress to produce corpora of texts annotated rhetorically. We believe this is also essential for multimodal documents. However, as we have detailed above, there are also significant issues that need now to be faced when we move away from linear presentations even to two-dimensional page-based presentations.

More practically, there are issues concerning how much information can be obtained from existing annotation and industry-standard markups: for example, the information maintained in professional document preparations tools such as QuarkXpress or Adobe Framemaker, InDesign, etc. Providing conversion tools to the kinds of linguistically motivated corpus annotations described here would open up a huge area of data. The genre and design knowledge encoded implicitly in style sheets and templates needs also to be made available so that it may be subjected to the kinds of study described above.

Of particular interest to us at present are further extensions across languages so as to compare cultural variation in visual/verbal presentations and further, more detailed comparison of documents variants created by repurposing (e.g., print-to-web, web-to-palmtop, etc.). In both cases, we are concerned that quite ordinary, everyday documents be considered equally, such as bills, consumer letters, instruction manuals, newspapers—these are the documents which users encounter in their everyday lives and understanding how they can be best structured could have significant practical benefits. The acquisition of annotated data across genre and cultures should also therefore be a high priority task.

Finally, we also require that the GeM annotation should

be able to fit into broader annotation schemes. Thus any kind of artifact that includes two-dimensional presentations (for example, a video embedded in a webpage) may also receive a GeM annotation for that component of the information offering. Our claims concerning coherence and consistency of information presentation decisions across text, visuals and layout can then be investigated here also. In such cases, the GeM annotation offers an annotation slice consisting of several annotation levels contributing to more comprehensive annotations that take in other important aspects of the artifact's design beyond that considered within the GeM model. In this respect, we consider it a crucial design feature that such annotation slices be additive and open rather than excluding and closed.

6. References

Elisabeth André. 1995. *Ein planbasierter Ansatz zur Generierung multimedialer Präsentationen*, volume 108. Infix, St. Augustin.

Roland Barthes. 1977. *Image – Music – Text*. Hill and Wang, New York.

John A. Bateman, Thomas Kamps, Jörg Kleinz, and Klaus Reichenberger. 2001. Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system. *Computational Linguistics*, 27(3):409–449.

Stephen Bernhardt. 1985. Text structure and graphic design: the visible design. In James D. Benson and William S. Greaves, editors, *Systemic Perspectives on Discourse, Volume 1*, pages 18–38. Ablex, Norwood, New Jersey.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27:3–43.

Nadjet Bouayad-Agha. 1999. Annotating a corpus with layout. In Richard Power and Donia Scott, editors, *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding or Retrieval of Documents*, pages 58–61, Cape Cod, Massachusetts, November. American Association for Artificial Intelligence.

Nadjet Bouayad-Agha. 2000. Layout annotation in a corpus of patient information leaflets. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000*, Athens, Greece. European Language Resources Association (ELRA).

Marc Corio and Guy Lapalme. 1998. Integrated generation of graphics and text: a corpus study. In M. T. Maybury and J. Pustejovsky, editors, *Proceedings of the COLING-ACL Workshop on Content Visualization and Intermedia Representations (CVIR'98)*, pages 63–68, Montréal, August.

Judy L. Delin and John A. Bateman. 2002. Describing and critiquing multimodal documents. *Document Design*, 3(2). Amsterdam: John Benjamins.

Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

Renate Henschel. 2002. GeM annotation manual. Gem project report, University of Bremen and University of Stirling, Bremen and Stirling. Available at http://purl.org/net/gem.

Gunther Kress and Theo Van Leeuwen. 2001. *Multimodal discourse: the modes and media of contemporary communication*. Arnold, London.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Ivandré Paraboni and Kees van Deemter. 2002. Towards the generation of document deixis reference. In Kees van Deemter and Rodger Kibble, editors, *Information sharing: reference and presupposition in language generation and interpretation*, pages 333–358. CSLI.

Klaus Reichenberger, Klaas Jan Rondhuis, Jörg Kleinz, and John A. Bateman. 1995. Effective presentation of information through page layout: a linguistically-based approach. In *Proceedings of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction*, San Francisco, California. ACM.

Karen A. Schriver. 1996. *Dynamics in document design: creating texts for readers*. John Wiley and Sons, New York.

Michael Twyman. 1982. The graphic presentation of language. *Information Design Journal*, 3:2–22.

Robert Waller. 1987. *The typographical contribution to language: towards a model of typographic genres and their underlying structures*. Ph.D. thesis, Department of Typography and Graphic Communication, University of Reading, Reading, U.K.

Towards a roadmap for Human Language Technologies: Dutch-Flemish experience

Diana Binnenpoorte[1,2], Catia Cucchiarini[2,3], Elisabeth D'Halleweyn[3], Janienke Sturm[2] and Folkert de Vriend[2]

[1]Speech Processing Expertise Centre (SPEX), Nijmegen, the Netherlands
[2]Department of Language and Speech, University of Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
{D.Binnenpoorte, C.Cucchiarini, F.deVriend, Janienke.Sturm}@let.kun.nl
[3]Nederlandse Taalunie, The Hague, The Netherlands
EdHalleweyn@ntu.nl

Abstract

In this paper we describe how the project "Dutch Human Language Technologies Platform" has contributed to creating the preconditions for establishing a roadmap for Human Language Technologies in the Dutch speaking area. Our overview of the results obtained so far reveals that the goals of all four action lines have been achieved and that there are clear directions for how to proceed in the near future. We hope that our experiences will be useful to other countries that intend to start similar initiatives.

1. Introduction

Establishing a roadmap for Human Language Technologies for a given language requires that first a number of important basic elements be defined, such as:

1. what is minimally required to guarantee an adequate digital language infrastructure for that language?
2. what is the current situation of HLT in that language?
3. what needs to be done to guarantee that at least what is required be available?
4. how can 3 best be achieved ?
5. how can we guarantee that once an adequate HLT infrastructure is available, it also remains so?

It is exactly these questions that were at the core of the activities that in the last two years were carried out within the framework of the Dutch-Flemish project "Dutch Human Language Technologies Platform". The ultimate aim of this project is to further the development and secure the usability of an adequate digital language infrastructure for Dutch, which is required to maximise the outcome of future efforts and to guarantee progress in the field of HLT.

In this paper we will report on our approach and our experiences in carrying out the activities envisaged in this project, because we think that this information can contribute to the aim of this workshop: establishing a roadmap for Human Language Technologies for the next decade.

2. The Dutch HLT Platform: action plan

The plan to set up a Dutch HLT platform was launched by the Dutch Language Union (Nederlandse Taalunie, NTU) which is an intergovernmental organisation established in 1980 on the basis of the Language Union Treaty between Belgium and the Netherlands. The NTU has the mission of dealing with all issues related to strengthening the position of the Dutch language (see also www.taalunie.org). In addition to the NTU, the relevant Flemish and Dutch ministries and organisations are involved in the HLT Platform. The various organisations have their own aims and responsibilities and approach HLT accordingly. Together they provide a good coverage of the various perspectives from which HLT policy can be approached.

The rationale behind the Dutch HLT platform was not to create a new structure, but rather to co-ordinate the activities of existing structures. The platform is a flexible framework within which the various partners adjust their respective HLT agendas to each other's and decide whether to place new subjects on a common agenda. Initially, the Dutch HLT platform was set up for a period of five years (1999-2004).

To achieve the objectives mentioned above, an *Action plan for Dutch in language and speech technology* was defined, which encompasses various activities organised in four action lines:

2.1. Action line A: performing a 'market place' function

The main goals of this action line are to encourage co-operation between the parties involved (industry, academia and policy institutions), to raise awareness and give publicity to the results of HLT research so as to stimulate market take-up of these results.

2.2. Action line B: strengthening the digital language infrastructure

The aims of action line B are to define what the so-called BLARK (Basic LAnguage Resources Kit) for Dutch should contain and to carry out a survey to determine what is needed to complete this BLARK and what costs are associated with the development of the material needed. These efforts should result in a priority list with cost estimates which can serve as a policy guideline.

2.3. Action line C: working out standards and evaluation criteria

This action line is aimed at drawing up a set of standards and criteria for the evaluation of the basic materials contained in the BLARK and for the assessment of project results.

2.4. Action line D: developing a management, maintenance and distribution plan

The purpose of this action line is to define a blueprint for management (including intellectual property rights), maintenance, and distribution of HLT resources.

Soon after the HLT Platform was set up it was decided that survey (action line B) and evaluation (action line C) be carried out in an integrated way because the actual availability of a product is not determined merely by its existence, but depends heavily on the quality of the product itself.

In the remainder of this paper we analyse the results of each action line in detail and in the final section we consider how this work has paved the way to a roadmap for Dutch HLT.

3. Action line A: results

In setting up HLT projects such as the *Spoken Dutch Corpus* and *NL-Translex*, much time was invested in the search for the appropriate responsible (funding) bodies in the Netherlands and Flanders. Moreover, various studies had indicated that the fragmentation of responsibilities made it difficult to conduct a coherent policy and meant that the field lacked transparency for interested parties. For these reasons the NTU, as the coordinator of the HLT Platform, stimulated the creation of a network aimed at:

disseminating the results of research in the field of HLT;

bringing together demand and supply of knowledge, products and services;

stimulating co-operation between academia and industry in the field of HLT.

After only two years of activity the HLT Platform has already produced important results. The success of Action line A is also partly due to the fact that the NTU acts as the National Focal Point (NFP) in the HOPE (Human Language Technology Opportunity Promotion in Europe) project. HOPE is a multi-country, shared-cost accompanying measure project of the IST-Programme of the European Commission that aims at providing awareness, bridge-building and market-enabling services to boost opportunities for market take-up of the results of national and European HLT RTD. The key focus is on helping to accelerate the volume of HLT transfer from the research base to the market by creating communities of interest between the critical players in the development and value chain. The aims of HOPE clearly coincide with the aims of Action line A.

At the beginning of the HOPE project an extensive informational website on the HLT sector in The Netherlands and Flanders was established by the NTU. This website provides up-to-date information on all relevant actors in the field of HLT (i.e. researchers, developers, integrators, users and policy makers) on how the HLT sector evolves on a cross-border Dutch/Flemish level, and on HLT related events throughout Europe. All this information is presented in Dutch and English.

The site also includes a calender of HLT events and a form for people who want to be included in the contacts database, as well as links to the HLTCentral website. All information on HLT related programmes and actions of the European Commission is provided on a separate website, established and maintained by subcontractor Senter/EG-Liaison, which is the most knowledgeable party on this subject. These two sites have one entry point from the HOPE point-of-view, via an intermediate site that was developed to provide clarity on where to find which information. This intermediate site (also in Dutch and English) has been placed on http://www.hltcentral.org/euromap/ and should be considered as the common homepage for the two websites. Visitors who do not find answers to their questions on the website can contact the NTU or Senter/EG-Liaison directly (preferably by e-mail) and may expect to receive quick and accurate replies.

Part of the infodesk task is also to conduct mailings to national contacts. These mailings are done on an ad-hoc basis, either at a third party's request (e.g. if an organizing committee wants to announce an event) or on the NFP's own initiative (e.g. if there is important news about an EC programme). From the beginning of the HOPE project, an extensive contacts database has been compiled by the NFP. At present, this database contains almost a thousand persons from over six hundred organisations in The Netherlands and Flanders. It is a valuable backbone for all information activities of the NFP.

The Dutch/Flemish NFP also visits companies with HLT related needs to demonstrate the benefits of HLT, to solicit a clear picture of the company's knowledge state and future plans, and to provide information of cross-linking services where appropriate. The NFP, in collaboration with its partners in The Netherlands and Flanders, has organised various seminars and workshops, which were attended by people from industry, academia, and policy institutions. The aim of such events is to further enhance awareness of recent developments in the HLT sector at the national and international level, such as the dissemination of information on European Commission HLT actions and their relevance to the national situation. Note that the cross-border Flemish/Dutch level should be considered here as the "national" level. The first national seminar took place in March 2001, and was a major event with over 150 participants. The second seminar was held in November 2001 and was directly related to the general survey carried out under action line B and C. Two other events are being organised for 2002. To conclude, we can safely state that in two years time the activities carried out within Action line A have certainly contributed to creating transparency and structure in the HLT field in The Netherlands and Flanders.

4. Results of Action lines B and C

The field survey comprised the following three stages: defining the BLARK for Dutch, making an inventory of available HLT resources, establishing a priority list. These three stages are described in more detail below.

4.1. Defining the BLARK

In defining the BLARK a distinction was made between applications, modules, and data:

Applications: refers to classes of applications that make use of HLT. The following classes were defined: CALL (Computer Assisted Language Learning), access control, speech input, speech output, dialogue systems, document production, information access, and multilingual applications or translation modules.

Modules: refers to the basic software components that are essential for developing HLT applications.
Data: refers to data sets and electronic descriptions that are used to build, improve, or evaluate modules.

In order to guarantee that the survey is complete, unbiased and uniform, a matrix was drawn up by the steering committee describing (1) which modules are required for which applications, (2) which data are required for which modules, and (3) what the relative importance is of the modules and data. This matrix (subdivided in language technology and speech technology) is depicted in Table 1, where "+" means important and "++" means very important.

This matrix serves as the basis for defining the BLARK. Table 1 shows for instance that monolingual lexicons and annotated corpora are required for the development of a wide range of modules; these should therefore be included in the BLARK. Furthermore, semantic analysis, syntactic analysis, and text pre-processing (for language technology) and speech recognition, speech synthesis, and prosody prediction (for speech technology) serve a large number of applications and should therefore be part of the BLARK, as well. Note that only language specific modules and data were considered in this survey.

Based on the data in the matrix the BLARK for Dutch should consist of the following components:

4.1.1. Language technology BLARK

Modules:
- Robust modular text pre-processing (tokenisation and named entity recognition),
- Morphological analysis and morpho-syntactic disambiguation,
- Syntactic analysis,
- Semantic analysis.

Data:
- Monolingual lexicon,
- Annotated corpus written Dutch (a treebank with syntactic, morphological, and semantic structures),
- Benchmarks for evaluation.

4.1.2. Speech technology BLARK

Modules:
- Automatic speech recognition (including tools for robust speech recognition, recognition of non-natives, adaptation, and prosody recognition),
- Speech synthesis (including tools for unit selection),
- Tools for calculating confidence measures,
- Tools for identification (speaker identification as well as language and dialect identification),
- Tools for (semi-) automatic annotation of speech corpora.

Data:
- Speech corpora for specific applications, such as CALL, directory assistance, etc.,
- Multi-modal speech corpora,
- Multi-media speech corpora,
- Multi-lingual speech corpora,
- Benchmarks for evaluation.

4.2. Inventory and evaluation

In the second stage, an inventory was made to establish which of the components - modules and data - that make up the BLARK are already available; i.e. which modules and data can be bought or are freely obtainable for example by open source. Besides being available, the components should also be (re-)usable. Obviously, components can only be considered usable if they are of sufficient quality; therefore, a formal evaluation of the quality of all modules and data is indispensable. Given the limited amount of time, only a formal evaluation was carried out by using a checklist with the following items: availability, programming code, platform, documentation compatibility with standard packages, reusability, adaptability and extendibility.
The information on availability, the matrix in Table 1 and the preliminary inventory were submitted to a group of HLT experts from both industry and academia, so that a balanced picture could be obtained.

Based on this information a second matrix was filled in which the availability of the components in the BLARK (cf. Table 2) was described. Availability in this matrix is expressed in numbers from 1 (‘module or data set is unavailable’) to 10 (‘module or data set is easily obtainable’).

At the end of the second stage, all information gathered was incorporated in a report containing the BLARK, the availability figures together with a detailed overview of available HLT resources for Dutch, a priority list of components that need to be developed, and a number of recommendations. This report was considered as being provisional as feedback on this version from a lot of actors in the field was considered desirable.

4.3. Feedback

One of the aims of Action lines B and C was that the majority of the actors in the HLT field would subscribe to the priorities and recommendations for the future. To this end, the provisional report containing the inventory, the priority lists and the recommendations was sent to a total of about 2000 people active in the HLT field who were asked to send their comments by email. After the relevant comments had been incorporated in the report, the same group of people was invited to participate in a workshop in which the results (overview, BLARK, priority lists and recommendations) were officially presented to the public.

On this occasion some people were given the opportunity to publicly present their views on the results of the survey. The workshop was concluded with a general discussion between the audience and a panel of five experts that were responsible for the survey.

The workshop provided useful information that could be used to complete the final report. A number of important points that emerged form this workshop are listed below:
- Cooperation between universities, research institutes and companies should be stimulated.
- For all components in the BLARK it should be clear how they can be integrated with off-the-shelf software packages. Furthermore, documentation and information about performance should be readily available.
- Control and maintenance of all modules and data sets in the BLARK should be guaranteed.
- Feedback from users on the quality and the performance of the various components should be processed in a structured way.

Special attention should be paid to the issue of open source policy and its possible effects for companies.

Modules	Data									Applications							
	mono lex	multi lex	thes	ann corp	unann corp	speech corp	multi ling	multi mod	multi media	CALL	access control	speech input	speech output	dialog system	doc prod	info access	transla-tion
Language Technology																	
Grapheme-phoneme conv.	++			++						+			++	++	+	+	
Token detection	++			+	++					+		+		+	+	+	+
Sent boundary detection	+			++	++					+		++	++	+	++	++	++
Name recognition	+	+	+	++	++	++				+		++	++	+	++	++	++
Spelling correction										+							
Lemmatizing	++			++	+					+		+	+	+	+	+	+
Morphological analysis	++			++	+					+		+	++	+	++	++	++
Morphological synthesis	++			++	+					+			++	+	++		++
Word sort disambig.	++			++	+					+		++	+	++	++	++	++
Parsers and grammars	++			++						+		++	++	++	++	++	++
Shallow parsing	++			++	++					+		++	++	++	++	++	++
Constituent recognition	++			++	+					+		++	++	++	++	++	++
Semantic analysis	++		++	++				++	++	+		++	++	++		++	++
Referent resolution	+		++	++	+					+		++		++	++	++	++
Word meaning disambig.	+		++	++	+					+		++	+	+	+	++	++
Pragmatic analysis	+		+	++				++	++	+		++	++	++		+	++
Text generation	++		++	++				++	++	+			++	++	++		++
Lang. dep. translation		++	++	++			++			+						++	++
Speech Technology																	
Complete speech recog.	++	+		++	+	++	+	++	++	++	++	++		++	++	++	++
Acoustic models	++	+		++	+	++	+	+	+	++	+	++		++	+	+	+
Language models	+			++	+	+	+	+	+	++	+	++		++	++	++	++
Pronunciation lexicon	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Robust speech recognition	+			+	+	+	+	+	++	+	+	++		++	+	+	+
Non-native speech recog.	+	++		+		++	++	+	+	++	+	+		+		+	+
Speaker adaptation	+			+	+	++	+	+	++	+	+	++		+	+	++	+
Lexicon adaptation	++	+		+		++	+	+	+	++	+	++	+	++	+	++	++
Prosody recognition	+	+		++	+	++	+	+	+	++	+	++		++	++	++	++
Complete speech synth.	++	+		+		+		+		+			++	++	+	+	++
Allophone synthesis	+	+		+		+		+		+			+		+	+	+
Di-phone synthesis	++	+		+		+		+		+			++	++	+	+	+
Unit selection	++	+		+		+		+		+			++	++	+	+	+
Prosody prediction for Text-to-Speech	++	+		+		+		+	+	++			++	++		+	++
Aut. phon. transcription	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Aut. phon. segmentation	++	++		+	+	++	+	+	+	++	+	+	+	+	+	+	+
Phoneme alignment	+	+		+		++	+	+	+	++	+	+		+			+
Distance calc. phonemes	+	+		+		++	+	+	+	++	+	+		+			+
Speaker identification	+			++	++	++	+	++	+	+	++	+		+		+	+
Speaker verification	+			++	++	++	+	++		+	++	+		+		+	+
Speaker tracking	+			++		++			++	+	++	+		+	+	+	+
Language identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Dialect identification	+	++		+	+	++	++	+	+	+	+	+		+		+	+
Confidence measures	+			+	+	++	+	++	+	++	++	++		++	+	+	+
Utterance verification	+			+	+	++	+	+	+	+	+	++		++	+	+	+

Table 1 *Overview of the importance of data for modules and the importance of modules for applications.*

Modules	Availability
Grapheme-phoneme conversion	8
Token detection	9
Sentence boundary detection	3
Name recognition	4
Spelling correction	3
Lemmatizing	9
Morphological analysis	
Morphological synthesis	
Word sort disambiguation	7
Parsers and grammars	3
Shallow parsing	2
Constituent recognition	5
Semantic analysis	3
Referent resolution	2
Word meaning disambiguation	2
Pragmatic analysis	1
Text generation	3
Language dependent translation	3
Complete speech recognition	4
Acoustic models	8
Language models	3
Pronunciation lexicon	5
Robust speech recognition	2
Non-native speech recognition	2
Speaker adaptation	2
Lexicon adaptation	2
Prosody recognition	2
Complete speech synthesis	6
Allophone synthesis	7
Di-phone synthesis	6
Unit selection	1
Prosody prediction for Text-to-Speech	3
Autom. phonetic transcription	3
Autom. phonetic segmentation	5
Phoneme alignment	8
Distance calculation of phonemes	8
Speaker identification	2
Speaker verification	2
Speaker tracking	2
Language identification	2
Dialect identification	2
Confidence measures	2
Utterance verification	2
Data	
Unannotated corpora	9
Annotated corpora	5
Speech corpora	4
Multi lingual corpora	3
Multi modal corpora	1
Multi media corpora	1
Test corpora	1
Monolingual lexicons	8
Multilingual lexicons	6
Thesaurus	4

Table 2 Availability of modules and data

4.4. Inventory, priority list and recommendations

The survey of Dutch and Flemish HLT resources resulted in an extensive overview of the present state of HLT for the Dutch language. By combining the BLARK with the inventory of components that are available and of sufficient quality, the following priority for language and speech technology lists were drawn up.

4.4.1. Priority list for language technology:

1. Annotated corpus written Dutch: a treebank with syntactic and morphological structures,
2. Syntactic analysis: robust recognition of sentence structure in texts,
3. Robust text-preprocessing: tokenisation and named entity recognition,
4. Semantic annotations for the treebank mentioned above,
5. Translation equivalents,
6. Benchmarks for evaluation.

4.4.2. Priority list for speech technology:

1. Automatic speech recognition (including modules for non-native speech recognition, robust speech recognition, adaptation, and prosody recognition),
2. Speech corpora for specific applications (e.g. directory assistance, CALL),
3. Multi-media speech corpora (speech corpora that also contain information from other media such as newspapers, WWW, etc.),
4. Tools for (semi-) automatic transcription of speech data,
5. Speech synthesis (including tools for unit selection),
6. Benchmarks for evaluation.

On the basis of the inventory and the reactions from the field the following recommendations were made:

- existing parts of the BLARK should be collected, documented and maintained by a central institution;
- the BLARK should be completed by financing the development of the resources prioritised;
- the BLARK should be made available to industry and academia through open source development;
- benchmarks, test corpora, and methods for evaluation and validation should be developed.
- the training of qualified HLT researchers should be encouraged.

5. Results of Action line D: the HLT Blueprint

In many cases official bodies such as ministries and research organisations are prepared to finance the development of language resources and no longer feel responsible for what should happen to these materials once the project has finished. However, materials that are not maintained quickly lose value. Moreover, unclear intellectual property right arrangements can create difficulties for exploitation. The purpose of action line D was to draw up a blueprint for management, maintenance and distribution of basic language materials that have been developed with government money. This includes, among other things, dealing with intellectual property rights issues, with the acquisition of resources, the adaptation of data and modules to other systems and applications,

making documentation available, providing a help desk function, maintaining and updating the material. Finally, this blueprint should provide guidelines for organizing a structural form of co-operation in this respect and should serve as an instrument for field organisations as well as for funding bodies.

The *Blueprint for management, maintenance and distribution of digital materials developed with public money (Blueprint)*, P. van der Kamp, T. Kruyt en P.G.J. van Sterkenburg) was prepared in the period 2000 -2001 by a team of language technology experts of the Institute for Dutch Lexicology, INL. In addition to the general aim of providing guidelines for the acquisition, management, maintenance and distribution of HLT materials, the *Blueprint* aims at providing information to be used by policy organisations when assessing research projects aimd at developing HLT materials, for preparing policy plans concerning the acquisition, management, maintenance and distribution of HLT materials and practical information on how to acquire, manage, maintain and distribute HLT materials, answers to questions concerning the (re)usability of HLT materials after the consortia that were set up for their development cease to exist. All this information is presented in the *Blueprint* in nine chapters that, apart from the introductory chapter 1, deal with the following topics:

- Acquisition of HLT resources (Chapter 2)
- Processing of acquired data (Chapter 3)
- Linguistic processing of HLT resources (Chapter 4)
- Management of HLT resources (Chapter 5)
- Maintenance of HLT resources (Chapter 6)
- Distribution of HLT resources (Chapter 7)
- Support to users (Chapter 8)
- Recommendations for future policy (Chapter 9)

The following eight recommendations for future policy are made in the final chapter:

1. An HLT agency is necessary
   In order to prevent that HLT materials developed with government money outside a permanent infrastructure become obsolete and therefore useless, a legal body such as an HLT agency is required.
2. Organisation form of HLT agency and role of NTU
   This HLT agency could be a Dutch-Flemish consortium of institutions and should not be related to one existing institution in particular, because not all expertise is available in one single institution. A co-ordinator could be appointed by NTU to ensure that the interests of the whole HLT field are represented.
3. Tasks of the HLT agency.
   Primary tasks of an HLT agency:
   Task 1. Management
   Task 2. Guarantee accessibility of data and software
   Task 3. Maintenance
   Secondary tasks of an HLT agency:
   Task 4. User support
   Task 5. Acquisition
   Distribution should be entrusted ELDA and LDC.
4. Costs to be met by the government.
   Since extra costs for personnel and hardware will be incurred, additional government funding is required.
5. Costs to be met by the users of the HLT agency
   Depending on the specific use and user, general conditions must be agreed on that guarantee fair tariffs.
6. Acceptance of HLT data and software by the HLT agency.
   The HLT agency can refuse HLT resources that do not meet certain quality standards or that are not essential for a wide range of applications.
7. International participation.
   The HLT agency should be given the possibility, through government funding, to participate in European and/or global projects that are related to its tasks.
8. Development and maintenance of HLT expertise.
   Given the considerable shortage of language and speech technologists, the government should stimulate policies that are aimed at developing and maintaining expertise in the field of HLT.

6. Future prospects

In the previous sections we have provided an overview of the results obtained within Action lines A and D. This has revealed that the aims identified in the *Action plan for Dutch in language and speech technology* have been achieved, at least for these two action lines. Now it remains to be seen how these results will be used in the future in order to achieve the ultimate aim of the "Dutch Human Language Technologies Platform" project: to further the development and secure the usability of an adequate digital language infrastructure for Dutch. To this end in the following sections we consider our future plans with respect to Action lines A (5.1) and D. (5.2).

6.1. Action line A

Since Action line A has already contributed to creating a co-operative framework in the HLT field in The Netherlands and Flanders, our future activities will be directed to maintaining and enlarging it. This entails among, other things, keeping our databases and websites up to date, ensuring communication between interested partners, gradually enlarging the initial network, identifying and promoting the inclusion of new representatives; increasing the visibility and the strategic impact of relevant results and new initiatives; fostering cooperation; providing a forum for discussing, exchanging and sharing experiences, best practices, information data and tools.

6.2. Action lines B and C: HLT priorities

The future activities of these two action lines will be directed to ensuring that the priorities identified in the survey are realized so that an adequate HLT infrastructure for Dutch is obtained.

6.3. Action line D: implementation of the recommendations in the HLT Blueprint

In the near future a number of Dutch-Flemish digital HLT resources will become available. These development projects, in many cases, do not provide a permanent infrastructure. As projects aimed at the development of digital basic resources mostly result in intermediary products, extra efforts and investments are needed in order to implement them in applications that find their way to the end users. Furthermore, when planning such large scale projects a lot of time is invested in building the

necessary structures (often at a supra-institutional level) and finding the right experts. The completion of a project often means that the managerial and operational structures cease to exist. Therefore it is of vital importance that the right measures are timely taken in order to ensure that the resources are stored in such a way that they will be expertly managed and maintained. When establishing an adequate infrastructure for maintenance of digital basic resources, proper attention should be given to a) intellectual rights, overall responsibility and co-ordination, b) actual physical management and maintenance of the resources and c) maintenance of expertise. In the following sections we will describe the facilities that we envisage to implement in the Dutch speaking area in the near future.

6.3.1. Necessary facilities

A. Intellectual rights, responsibility, co-ordination: ***NTU***

A careful transfer of intellectual rights is of crucial importance to the exploitation of resources. Furthermore, after completion of projects a visible policy responsibility is needed, even if the actual management and maintenance is carried out by an HLT agency (see B).

Organisational structure: The NTU (Nederlandse Taalunie/Dutch Language Union), representing a permanent Dutch-Flemish infrastructure, can act as the appropriate legal body handling all legal affairs. A member of the NTU will be appointed as co-ordinator and supervise from a policy point of view management, maintenance and exploitation of HLT basic resources that are contributed to the HLT agency (see B)..

The NTU will look after the interests of the entire HLT field and will function as a kind of 'broker' by:

- supervising the activities of the HLT agency (see B) and the various HLT committees (see C);
- looking after legal issues;
- stimulating the application of international standards;
- stimulating funding bodies to stipulate that in proposals proper attention is paid to allocating funding for management and maintenance and that resources financed with public funding be made available through the HLT agency;
- playing an intermediate role in the acquisition of digital data, e.g. from the industry.

B. Management and maintenance of digital resources: HLT agency

The *Blueprint* recommends the co-operation of the institutes in a consortium, an **HLT agency**, as this makes it possible to use dispersed expertise and infrastructure. This construction clearly has a number of advantages:

- efficient use of persons and means can be cost-reducing;
- combining resources and bringing together different kinds of expertise can create surplus value (e.g. extra applications);
- offering resources through one window (one-stop-shop) will create optimal visibility and accessibility;
- in international projects the Dutch language area can act as a strong partner;

Organisational structure: The HLT agency can take the form of a Dutch-Flemish consortium of organisations contributing their resources and expertise in a virtual resource centre. These organisations should strike binding agreements for a determined period of time. One Dutch-Flemish organisation (e.g. the Dutch Institute of Lexicology in Leiden) should be appointed as responsible co-ordinator.

- management: taking the appropriate (mostly technical) measures so as to make sure that data and software remain operational and usable;
- accessibility data and software: facilitating reusability of HLT resources: e.g. technical, legal and administrative settlements so as to optimise the route from developer via HLT agency to the distributor;
- maintenance: taking the appropriate measures to ensure long-term usability of data and software: technical maintenance of formats of HLT data, HLT software, system and application software, equipment; maintenance of legal contracts; content management of the HLT data and annotations;
- service: help desk, service to the users of the HLT data and HLT software (e.g. advising, maintenance of website and mailing lists, supplying tailor made data or software on demand);
- acquisition: active acquisition of HLT data and HLT software developed by the industry or research institutes;
- evaluation and validation: contributing to establishing international standards and methods for evaluating and validating HLT resources.

For the actual, physical distribution of the resources appeal will be made on the expertise of organisations s.a ELRA and LDC as they have the proper expertise and marketing tools.

C. Expertise: Dutch-Flemish steering committees and HLT management committee

In dissolving the managerial and operational infrastructure after the completion of a project, valuable specific knowledge concerning the project may be lost causing difficulties in the exploitation of the results. All the same it would not be realistic to maintain these structures. A solution would be to install a number of Dutch-Flemish **steering committees** and one co-ordinating Dutch-Flemish **HLT management committee**. The tasks of these committees should not be too heavy, but to ensure continuity and effectiveness a strong secretarial support should be provided

Organisational structure: For each completed large scale project the results of which are contributed to the HLT agency, a steering committee should be installed. Each steering committee delegates one representative to a co-ordinating HLT management committee. For small scale projects it has to be examined whether the necessary expertise is already present in the HLT management committee. Probably one expert, responsible for the combined 'small' projects, will be added to this committee. The various committees should receive the appropriate secretarial support.

Tasks: The steering committees will be responsible for specific resources and specific domains. They will

- act as a knowledge base for questions concerning the resources contributed to the HLT agency;
- act as intrinsic supervisors on management, maintenance and exploitation of specific resources;
- act as advisors in specific domains s.a. language and speech technology, terminology, lexicology;
- be instrumental in the organisation of 'major repairs' of the resources that are put in their custody;
- be instrumental in developing the appropriate infrastructure for new projects or updating of existing results in their domain.

The HLT management committee will be responsible for the co-ordination, overall management, maintenance and distribution of HLT resources. It will

- act as general knowledge base and give advise in the broad field of language and speech technology, terminology, lexicology etc..
- act as general intrinsic supervisor on management, maintenance and exploitation of finished resources;
- be instrumental in developing the appropriate personnel infrastructure for new projects or updating of existing results.

6.3.2. Financing

Since the exploitation of basic resources will not result in considerable revenues, the authorities have expressed their explicit wish to make these resources available as broadly as possible. This results in keen prices: cost price for non-commercial research, a higher but not prohibitive price for commercial organisations. Consequently, the implementation of the above mentioned structures requires extra funding. Since a considerable percentage of the development costs should be allocated to management and maintenance, by combining the infrastructures required for different projects the percentage the costs would decrease. This applies as much to the material infrastructure (equipment, data, software, licences, etc…) as to the immaterial infrastructure (experts, personnel etc.). As is stressed in the recommendations of the *Blueprint*, the activities of the HLT agency cannot be carried out by the consortium partners in addition to their daily work, but require extra staff. Based on the data in the *Blueprint* and on experiences in other projects, a number of persons will be appointed at one or more of the organisations forming the HLT agency (e.g. experts on language and speech technology, IT-specialist, administrative personnel etc.). One overall co-ordinator and at least one secretary of the committees will be appointed at the NTU.

It is to be expected that the costs will increase with the increase of project results contributed to the HLT agency. These costs should be covered with funds allocated to management, maintenance and accessibility at the start of the development of new projects.

6.3.3. Conclusions

After the completion of projects aimed at developing HLT resources, efforts are needed to ensure long-term usability of the results. Timely attention to intellectual property rights, management, maintenance and distribution can guarantee that investments pay off in the future. In this respect, it is recommended, to make optimal use of existing expertise and infrastructure. In concrete this would mean that in the Dutch speaking area:

- the co-ordinating policy responsibility and as much intellectual property rights as possible should be placed in the hands of the NTU;
- the actual exploitation (management, maintenance and distribution) should be entrusted to a Dutch-Flemish HLT agency, that will take the shape of a consortium of institutions but acts as a one-stop-shop of digital HLT resources for the Dutch language
- the existing expertise should be combined as much as possible in a number of Dutch-Flemish steering committees consisting of representatives of projects, the results of which are contributed to the HLT agency and a co-ordinating Dutch-Flemish HLT management committee.

The NTU envisages to implement the above mentioned structures in its new long-term policy plan (2003-2007).

7. General conclusions

In this paper we have reported on the activities that in the last two years have been carried out within the framework of the project "Dutch Human Language Technologies Platform". In particular, we have focussed on two of the four action lines within this project: Action line A, which was aimed at raising awareness of the results of HLT research and promoting communication among interested partners, and Action line D which was concerned with management, maintenance and distribution of HLT resources.

Our overview of the results obtained so far has revealed that a cooperative framework has been created and that there are clear plans to set up a structure that will take care of all HLT resources developed with public funding, so that they will remain available for all interested parties: an HLT agency. In other words, the goals of action lines A and D have been achieved (for the results of B and C, the reader is referred to Binnenpoorte et al. (2002)) and clear directions for how to proceed in the near future have also been outlined. To conclude, it seems that in the Dutch speaking area pioneering work has been carried out from which other countries can probably profit in their attempts to start similar initiatives.

8. Acknowledgements

We are indebted to the steering committees of Action lines B, C, and D and to the authors of the Blueprint (P. Van der Kamp, T. Kruyt, and P. Van Sterkenburg) and of the Report B and C (G. Bouma, W. Daelemans, A. Dirksen, D. Heijlen, F. de Jong, J.-P. Martens, A. Nijholt, H. Strik, D. van Compernolle, F. van Eynde, and R. Veldhuis) for their invaluable contribution to the work presented in this paper.

9. References

Binnenpoorte, D., de Vriend, F., Sturm, J., Daelemans, W., Strik, H., and Cucchiarini, C. (2002). A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch. In *Proceedings of LREC2002*.

ANNEXES

SPEECH-RELATED TECHNOLOGIES

Where will the field go in 10 years?

Niels Ole Bernsen, NISLab, Denmark (editor)

Abstract

This paper is a draft position paper for discussion at the ELSNET Brainstorming Workshop 2000-2010 in Katwijk aan Zee, The Netherlands, on 23-24 November, 2000. The paper first describes some general emerging trends which are expected to deeply affect, or even transform, the field of speech technology research in the future, including trends towards advanced systems research, natural interactivity, multimodality, and medium-scale science. A timeline survey of future speech-related technologies is then presented followed by analysis of some of the implications of the proposed timelines. Timeline projections may turn out to have been false, of course, but even their turning out to be true is subject to future actions which are (not) taken to make them true. Accordingly, the final part of the paper discusses some actions which would seem desirable from the point of view of strengthening the position of European speech-related research.

1. Introduction

The term *speech-related research* has been chosen to designate the topic of the present paper for lack of ability to invent a more appropriate term, if there is one. At least, the term partly manages to convey the author's expectation that the field of speech research will change rather dramatically in the coming ten years as speech technologies become merged with other technologies into a field which, so far, lacks a name.

According to many observers, the coming decade will be the decade of speech technologies. Computer systems, whether stationary or mobile, wired or wireless, will increasingly offer users the opportunity to interact with information and people through speech. This has been made possible by the arrival of relatively robust, speaker-independent, spontaneous (or continuous) spoken dialogue systems in the late 1990s as well as through the constantly falling costs of computer speed, bandwidth, storage, and component miniaturisation. The presence of a speech recogniser in most appliances combined with distributed speech processing technologies will enable users to speak their native tongue when interacting with computer systems for a very large number of purposes. Although no doubt exaggerated as just presented, there probably is some truth to this vision of a breakthrough in the application of speech technologies in the coming years. If this is the case, it would seem worthwhile that we lift our sights and take a long-term view of the issues ahead. This may help setting a reasonable research agenda for the coming years of advanced speech systems research and development, one which does not succumb to the usual hype associated with fashionable technologies. Today, some believe that "the speech problem" has been solved already. Some believe that speech, because of its naturalness, is the solution to every conceivable problem of user-system interaction. On the other hand, surprising as it may seem, some human factors and interactive systems experts believe that we have just arrived at the touch-tone telephony stage and share no notion of the actual state-of-the-art in the field with its practitioners. Since

all of those beliefs are far from the truth, it is important to provide a more balanced picture of the state-of-the-art in speech technologies in order to set the stage for solid progress.

In what follows, Section 2 presents some trends in the speech-related research field. Section 3 excels in guesswork by estimating the times of appearance of a range of novel speech-related technologies. Section 4 discusses implications of the timelines presented in Section 3. Section 5 proposes a series of actions which would appear appropriate given the preceding discussion.

2. Some Trends

The speech field is making progress on a broad scale as demonstrated by the 900 or so papers and posters presented at the recent International Conference on Spoken Language Processing (ICSLP) in Beijing, October 2000. [To be illustrated by listing topics.] Three points may be made on the preceding list of current topics in speech research. Firstly, the wealth of topics that are being addressed in current fundamental and applied research obviously demonstrates that "the speech problem" has *not* been solved but continues to pose a series of major research challenges. [Mention some of them.] Secondly, the breadth of the speech topics that are being addressed could be taken as evidence that the speech field is simply doing *business as usual,* albeit on a larger and more ambitious scale than ever before. Thirdly, however, it is clear from the topics list that *the speech field is no longer separate from many other fields* of research but is in a process of merging into something which might perhaps be called the general field of interactive technologies. This latter trend, it may be argued, is the single most important factor which will influence the speech field in the future and which already suggests that the field is in a state of profound transformation.

Interactive technologies

It is relatively straightforward to explain why the speech field is gradually merging into the general field of interactive technologies. Since speech now works for a broad range of application purposes, a rapidly growing fraction of the speech research community are becoming involved in advanced interactive *systems* research rather than continuing to work on improving the speech *components* which form part of those systems. In advanced interactive systems research, speech is increasingly being used not as a stand-alone interactive modality as in, e.g., spoken language dialogue systems over the telephone, speech dictation systems, or text-to-speech systems, but as a modality for exchanging information with computer systems in combination with other modalities of information representation and exchange. Moreover, speech is not just an interactive technology among many others. Spontaneous speech is an extremely powerful input/output modality for interacting with computer systems, a modality which, furthermore, is available and natural to the large majority of users without any need for training in using it for interactive purposes.

The ongoing shift from speech components research to research on integrating speech in complex interactive systems has a number of important implications for the speech field. Speech researchers are becoming systems researchers and engineers. Far more than components research, systems research and engineering is exposed to the full complexity of today's world of information and telecommunications technologies. Few, if any, groups can build full systems on their own from scratch. To stay competitive, they have to *follow closely* the global developments in relevant systems architectures, platforms, toolkits, available components of many different kinds, de facto standards, work in standards committees, market trends etc. They need larger and much more *interdisciplinary teams* in order to keep up with competitive developments. They need *access* to platforms and component technologies in order to avoid having to do everything by themselves. And they need expertise in *software systems engineering best practice as specialised to the kind of systems*

they are building, including expertise in systems and usability evaluation. As we shall see in Section 4, they need even more than this, such as *hardware* access or expertise, development *resources, behavioural research* in new domains, and skills in *form and contents design.*

Compared to traditional research on improving a particular speech component technology, the world of advanced interactive systems research would appear to be orders of magnitude more complex. Moreover, that world is quite diffuse for the time being. It does not have a single associated *research community,* being inhabited instead by researchers from most traditional ITC (Information Technologies and Telecommunications) research communities. The world of advanced interactive systems research does not have any clear *evolutionary direction,* being characterised rather through ever-changing terms of fashion, such as 'ubiquitous computing', 'things that think', 'wearable computing', 'the disappearing computer' or 'ambient intelligence'. Significantly, all or most of those terms tend to refer to combined hardware and software systems rather than to components, and none of them refer to the traditional communities in the ITC field, such as speech processing, natural language (text) processing, machine vision, robotics, computer graphics, neural networks, machine learning, or telecommunication networks. Indeed, most of our current stock of inspired and visionary terms for describing the future of interactive technologies tends to be rather vague with regard to the technologies which they include or, if any, exclude.

Rather than trying to clarify what might be meant by the terms of fashion mentioned above, it may be useful to look at two other developments in conceptualising the field of advanced interactive systems research of which speech research has begun to form a part. To be sure, the concepts to be discussed are expressed by fashion terms as well, but at least it would seem that those concepts are of a more systematic and theoretically stable nature at this point.

Natural interactivity

When being together, most humans interact through speech when they exchange information. The telephone allows them to use spoken interaction at a distance as well, and the function of the telephone will soon be shared, or even taken over, by computing systems. When humans interact through speech, it does not matter if they are just a twosome or if they are more than two together. Moreover, except when speaking over the telephone, speech is not their only modality for information exchange. Gesture, lip movements, facial expression, gaze, bodily posture, and object manipulation all contribute to adding information, however redundant, to the spoken message. Together with speech, those modalities constitute *full natural human-human communication.* Moving beyond current technologies, we envision not just a single human speaking on the telephone or to a (desktop) computer in order to get a particular task done. Rather, the vision is one in which multiple humans speak together whether or not they are in the same physical location whilst using the system as an increasingly equal partner in communication. The system mediates their communication when needed, understands full natural communication, and produces full natural communication itself, increasingly acting as its human counterparts in communication. In order to take this vision into account, it would seem timely to abandon the traditional model of interaction which is called 'human-computer interaction', and replace it with the more general model of *natural human-human-system interaction* (HHSI). Natural HHSI, it appears, it a necessary end-point of current research in speech technologies. Thus, natural interactivity may serve as an important, even if distant, guidepost for the role of speech research in the complex world of interactive systems research.

The received picture of the role of theory in engineering goes something like this. It is hardly ever possible to deduce from theory a complete specification of the artefact that would constitute an optimal solution to some engineering problem. The reason is that the complexity

of the problem space involved always exceeds the power of theory. On the other hand, without theory (of physics, chemistry, computation etc.), it would not have been possible to build many of the artefacts we use in our daily lives. Thus, theory has a necessary supporting function in engineering. This is clear in the case of natural interactivity. To achieve the ultimate goal of natural HHSI, we need far better theory than is available at present: about how humans behave during natural interaction, about the behavioural phenomena which are relevant to the development of fully natural interactive systems, about how these phenomena are interrelated, about how they should be encoded etc. We also need a novel theory of natural communication which can replace speech acts theory and discourse theory by taking the notion of a complete communicative act as its basic notion.

Multimodality

The trend towards multimodal interactive systems reflects the trend towards blending of traditional research communities noted above as well as the increasing role of speech in future interactive systems. Multimodal systems are systems which offer the user combinations of input/output modalities for (or ways of) exchanging information with computer systems. Given the naturalness and expressive power of speech, speech input and speech output have the potential for becoming key modalities in future interactive systems. However, compared to natural interactivity, our current understanding of multimodality is much less capable of providing guideposts for future advanced interactive systems research in general and research on multimodal systems which include speech modalities in particular. Much too little is known about how to create good modality combinations which include speech for a variety of interactive purposes. This topic has become an active field of research, however (Bernsen 1997a, Benoit et al. 2000, Bernsen 2001). Further progress in this field is likely to complement research on natural interactivity in providing guideposts for speech-related research in the complex world of advanced interactive systems. In fact, these two research directions are intertwined in so far as it remains an open issue for which application purposes technologies, such as, e.g., animated speaking characters might provide useful solutions.

Medium-scale science

The final trend to be mentioned is the trend towards medium-scale science in advanced interactive systems research. Increasingly, it is becoming evident that the standard 3/4/5-team, low-budget, 3-year isolated advanced systems research project is often an inefficient means of achieving significant research progress. In many projects, the participants share discouraging experiences, such as the following: even if small, the project is only able to start almost one year after its conception because of the administrative processing needed to release the funding for the project; when the project begins, the participants discover that their objectives have already been achieved elsewhere; the participants spend the first half of the project trying to identify the best platform to work from only to discover that they cannot get access to it; the participants spend half of the project building and putting together a low-quality version of the contextual technologies they need before they can start addressing their core research objectives; at the start of the project, the participants realise that it will take too long to produce the data resources they need, such as tagged corpora, and decide instead to work with sub-optimal resources which they can get for free; etc. One way to avoid, or reduce the number of, such experiences is to launch larger-scale concerted research efforts which have a better chance of moving beyond the state of the art. World-wide, experiments are currently underway on how to carry out such medium-scale science. In the US DARPA Communicator project which addresses spoken language and multimodal dialogue systems, for instance, all participants start from shared core technologies without having to build these themselves (http://fofoca.mitre. org/). In the German SmartKom project which addresses multimodal

communication systems, the budget is large enough for the participants to build and integrate the technologies needed (http://smartkom.dfki.de/start.html). In the European Intelligent Information Interfaces (i3, http://www.i3net.org/) and CLASS (http://www.class-tech.org/) initiatives, whilst the traditional 3-year small-scale project topology has been preserved, major efforts are being made to promote cross-project collaboration, synergy, and critical mass.

For reasons too obvious to mention, relatively small-scale research should continue to exist, of course. Still, the complexity of the world of advanced interactive systems research is not likely to go away. This raises the question of whether we need more medium-scale science and less small-scale science in order to make efficient use of the funds available for advanced interactive systems research. If this question is answered in the affirmative, the important issue becomes how best to do medium-scale science, i.e. which model(s) to adopt for the larger-scale research efforts to come.

3. Estimated Technology Timelines

This section attempts to estimate the time of first appearance of a broad selection of generic and/or landmark speech technologies including natural interactivity technologies and multimodal technologies involving speech. Some qualifications are necessary to the proper interpretation of the proposed predictions. Despite the numerous uncertainties involved in estimating technology progress, timelines, when properly estimated, qualified, and peer reviewed, do seem a useful means of conveying a field's expectations to the outside world and serving as a basis for actions to be undertaken to support research in the field.

Qualifications

(a) As in all timeline forecasts, there is some uncertainty in the forecasts below with respect to whether the technology is deployable or will in fact have been deployed in products at the suggested time. The claim for the figures below rather tend towards the *deployable* interpretation which is the one closest to the point of view of research. The *actual deployment* of a deployable technology is subject to an additional number of factors some of which are unpredictable, such as company technology exploitation strategies, pricing strategies, and the market forecasts at deployability time. Thus, several years may pass before some of the technologies below go from deployability to actually being used in mass products. This implies that one cannot from the estimations below construct scenarios for the Information Society in which people in general will be using the described technologies at the times indicated. In other words, the years below refer to "earliest opportunity" for actual deployment in what may be sometimes rather costly systems to be embraced by relatively few customers. Similarly, given the fact that there are thousands of languages in the world, it goes without saying that a technology has been established when it works in at least one of the top languages, a "top language" being defined as a language used by developers in the more affluent parts of the world.

(b) Another point related to (a) above is to do with underlying "production platforms". For many advanced, and still somewhat futuristic, speech and language -related systems, it is one thing to have produced a one-of-a-kind demonstrator system but quite another to have produced the system in a way which enables oneself or others to relatively quickly produce more-of-the-same systems in different application domains. An example is the so-called intelligent multimedia presentation systems which will be discussed in more detail in Section 4. Several examples exist, such as the German WIP system and corresponding systems from the USA. However, as long as we haven't solved the problem of how to produce this kind of system in a relatively quick and standardised way, intelligent multimedia presentation

systems are not going to be produced in numbers but will remain research landmarks. The timeline list below mostly avoids mentioning systems of this kind, assuming for the kinds of systems mentioned that the "production platform" issue has been solved to some reasonable extent at the time indicated.

(c) There is some, inevitable because of the brevity of the timeline entries, vagueness in what the described technologies can actually do.

(d) It is assumed that, after a certain point in time which could be, say, 2006, the distinction between technology use for the web and technology use for other purposes will have vanished.

(e) There is no assumption about *who* (which country, continent, etc.) will produce the described landmark results. However, given the virtually unlimited market opportunities for the technologies listed as a whole, it is expected that a consolidated technology timeline list will command keen interest among decision makers from industry and funding agencies.

(f) There is nothing about (software) agent technologies below. It is simply assumed that what is currently called software agent technologies will be needed to achieve the results described and will be available as needed.

(g) In principle, of course, any technology timeline list is subject to basic uncertainty due to the "if anything is done about it" –factor. If nothing will be done, nothing will happen, of course. However, most of the technologies listed below are being researched already and the rest will no doubt be investigated in due course. The uncertainty only attaches to who will get there first with respect to any given technology, who will produce the product winners, and how much effort will be invested in order to achieve those results before anybody else.

Technology timelines

Basic technologies

Hypotheses lattices, island parsing, spotting in all shapes and sizes for spoken dialogue	2001
Continuous speech recognisers in OSs for workstations in top languages	2002
Continuous speech recognisers in mobile devices (10000 words vocabulary) in top languages	2003
High quality competitive (with concatenated speech) formant speech synthesis in top languages	2003
Task-oriented spoken dialogue interpretation by plausibility in context and situation	2003
Generally usable cross-language text retrieval	2003
Multilingual authoring in limited domains by constructing conceptual representations	2003
Usable ontological lexicons for limited domains	2003
Usable translation systems for written dialogues (multilingual chatting)	2003
Useful speaker verification technology	2004
Seamless integration of spoken human/machine and human/human communication	2004
First on-line prosodic formant speech synthesis in top languages	2004
Simple task-oriented animated character spoken dialogue for the web	2004
Concept-to-speech synthesis	2004
Stylistically correct presentation of database content	2004
Superficial semantic processing based on ontological lexicons	2004

Max. 2000 words vocabulary task-oriented animated character dialogue for the web	2005
Prosodic formant speech synthesis replaces concatenated speech in top languages	2005
Full free linguistic generation (from concepts)	2005
Robust, general meta-communication for spoken dialogue systems	2005
Writer-independent handwriting recognition	2005
Learning at the semantic and dialogue levels in spoken dialogue systems	2006
Useful multiple-speaker meeting transcription systems	2006
Task-oriented fully natural animated characters (speech, lips, facial expression, gesture) output (only)	2007
Context sensitive summarization (responsive to user's specific needs)	2007
Answering questions by making logical inferences from database content	2007
Speech synthesis with several styles and emotions in top languages	2008
Continuous speech understanding in workstations with standard dictionaries (50000 words) in top languages	2008
Controlled languages with syntactic and semantic verification for specific domains	2008
Large coverage grammars with automatic acquisition for syntactic and semantic processing for limited applications	2008
Task-oriented fully natural speech, lips, facial expression, gesture input understanding and output generation	2010

Systems

First personalised spoken dialogue applications (book a personal service over the phone)	2002
Useful speech recognition-based language tutor	2003
Useful portable spoken sentence translation systems	2003
Useful broadcast transcription systems for information extraction	2003
First pro-active spoken dialogue with situation awareness	2003
Current spoken dialogue systems technology for the web (office, home)	2004
Satisfactory spoken car navigation systems	2004
Current spoken dialogue systems technology for the web (in cars)	2005
Useful special-purpose spoken sentence translation systems (portable, web etc.)	2005
High quality translation systems for limited domains with automatic acquisition	2005
Small-vocabulary (>1000 words) spoken conversational systems	2005
Medium-complexity (wrt. semantic items and their allowed combinations) task-oriented spoken dialogue systems	2005
Multiple-purpose personal assistants (spoken dialogue, animated characters)	2006
Task-oriented spoken translation systems for the web	2006
Useful speech summarisation systems in top languages	2006
Useful meeting summarisation systems	2008
Usable medium-vocabulary speech/text translation systems for all non-critical situations	2010
Medium-size vocabulary conversational systems	2010

Tools, platforms, infrastructure

Standard tool for cross-level, cross-modality coding of natural interactivity data	2002
Infrastructure for rapid porting of spoken dialogue systems to new domains	2003
Platform for generating intelligent multimedia presentation systems with spoken interaction	2005
Science-based general portability of spoken dialogue systems across domains and tasks	2006

Other problems which were strongly felt when producing the list above include: (i) the fact that there is plenty of continuity in technology development. "Continuity" may not be the right term because what happens is that what is later perceived as a new technological step forward is constituted by a large number of smaller steps none of which could be mentioned in a coarse-grained timeline exercise such as the one above. General speaker identification, robust speech recognition in hard-to-model noise conditions, "real" speaker-independent recognition (almost) no matter how badly people speak, or pronounce, some language, are all examples of minute-step progress. (ii) Another problem is to do with speech in fancy-termed circumstances, such as 'ambient intelligence' applications. It may be that there is a hard-core step of technological progress which is needed to achieve speech-related ambient intelligence but then again, may be there isn't. Maybe this is all a matter of using the timelined speech technologies above for a wide range of systems and purposes. Similarly, it is tempting to ask, for instance: "When will I have a speech-driven personal assistant?". But everything depends on what the personal assistant is supposed to be able to do. Some personal assistant technologies exist already. Thus, it does not seem possible to timeline the appearance of speech-driven personal assistants even if this might be attractive for the purpose of advertising the potential of speech technologies.

How well is Europe doing?

No attempt has been made, so far, to annotate the technology timelines with indications of how well, or how badly, European research is doing and hence how likely it is that a particular technology will be made deployable in Europe before anywhere else. In most of the timelined cases above, this would seem to depend primarily on the financial resources and research support mechanism which will be available to European research in the coming decade. In some cases, the US is presently ahead of Europe, such as with respect to continuous speech recognisers in workstations or broadcast transcription systems. In other cases, Europe has the lead, such as in building a standard tool for cross-level, cross-modality coding of natural interactivity data, continuous speech recognisers in mobile devices, advanced spoken dialogue systems, and spoken car navigation systems.

Beyond 2010

Beyond 2010 lie the dreams, such as unlimited-vocabulary spoken conversational systems, unlimited-vocabulary spoken translation systems, unlimited on-line generation of integrated natural speech, lips, facial expression and gesture communication, unlimited on-line understanding of natural speech, lips, facial expression and gesture communication by humans, summarisation-to-specification of any kind of communication, multimodal systems solutions on demand, and, of course, full natural interactive communication.

4. Implications of the Timelines

When analysing the implications of the timelines in Section 3, a number of uncertainties come up with respect to how the market for speech products will develop. At present, most speech

products are being marketed by some 5-10 major companies world-wide. These companies are growing fast as are hundreds of small start-up companies many of which use basic technologies from the larger technology providers. It may be assumed that this market structure will not continue in the future. Rather, speech recognition and synthesis technologies would seem likely to become cheap, or even free and open source, components which will come with all manner of software and hardware systems. The implication is that all ITC providers who want to, will provide value-added speech products and that the basic speech technologies will not be dominated by a small number of large suppliers. Some important share of the speech market, including de facto standards in various areas, will probably be picked up by large custom software and mobile phone technology suppliers, such as Microsoft and Nokia, but that is likely to happen in any realistic scenario for the coming decade. The conclusion is that, during the coming decade, speech will be everywhere, in all sorts of products made by all sorts of companies. But will speech be everywhere in bulk? This raises a second uncertainty.

In one scenario, speech will be present in all or most ITC products by 2010, and speech will be popular and will be used as much as input keys, input buttons, and output graphics displays are being used today. In another scenario, however, speech uptake will be slow and arduous. Several reasons could be given for the latter scenario. Thus, (a) it may take quite some time before speech recognition is being perceived by users to be sufficiently robust to make users switch to speech where speech is better ideally. (b) It may take quite some time before the field and the market has sorted out when to use speech as a stand-alone modality and when to use speech in combination with other input/output modalities. If these two (a + b) take-up curves do not grow in any steep manner, speech may still be widespread by 2010, but speech will still not be as important an input/output modality as it is likely to become later on. For the time being, we would appear to have too little information to be able to decide between the two scenarios just discussed. There is simply not enough data available on user uptake of speech technologies to enable a rational decision to be made.

Exploitation today

Already today, there is a great exploitation potential for speech technologies because of the simple facts that (i) the technologies which already exist in a few top languages could be ported to hundreds of other languages, and (ii) the types of applications which already exist can be instantiated into numerous other applications of similar complexity. At this end of the speech technology spectrum, the emphasis is on flexible and versatile production platforms, quality products, and low-cost production rather than on research. This is particularly true of low-complexity over-the-phone spoken language dialogue information systems using continuous speech input. Users would seem to have adopted these systems to a reasonable extent already. The same degree of user acceptance does not appear to characterise the uptake of, e.g., spoken language dictation systems or simple spoken command systems for operating screen menus. Even if purchased by widely different groups of users, the former would appear to be used primarily by professionals, such as lawyers and medical doctors, and the latter hardly seems to be used at all. Also, text-to-speech systems for the disabled and increasingly for all users, do appear to have a significant exploitation potential already.

Key technologies: speech-only

The timelines in Section 3 highlight a series of key speech-only technologies which are still at the research stage, including:

- prosody in on-line speech synthesis;
- multi-speaker broadcast and meeting transcription;

- speech summarisation;
- speech translation; and
- conversational spoken dialogue.

Prosody in on-line speech synthesis

Prosody in on-line speech synthesis is probably important to the speed of take-up of speech technologies because users would appear likely to prefer prosodic speech output to non-prosodic speech output. However, there do not seem to exist firm estimates as to how much prosody matters. Reasonably clear and intelligible non-prosodic text-to-speech already exists for some top languages and might turn out to be satisfactory for most applications in the short-to-medium term.

Multi-speaker broadcast and meeting transcription

Multi-speaker broadcast transcription forms the topic of massive US-initiated research at the moment and appears likely to start becoming widely used in practice relatively soon. Like *meeting transcription* technology, multi-speaker broadcast transcription technology has a large potential for practical application as well as for acting as a driving force in speech and natural language (text) processing research. Once multi-speaker broadcast speech audio and meeting speech audio can be useably transcribed so that first application paradigms for these technologies have been achieved, the transcriptions can be further processed by other technologies, such as speech summarisation and speech translation technologies. It would be very valuable for European speech research if Europe could launch a meeting transcription technology evaluation campaign before the US (evaluation campaigns will be discussed below).

Speech summarisation

Speech summarisation is being experimented with already, often by using text or transcribed speech instead of raw speech data. Speech and text summarisation technology including intelligent speech and text search would seem to hold enormous potential by enabling users to obtain at-a-glance information on the contents of large repositories of information. The same applies to related technologies, such as question-answer systems which enable the user to obtain answers to specific questions from large repositories of information. Progress in these fields is difficult because of the difficulty of the research which remains to be done. However, the difficulties ahead are counter-balanced by expectations that far-less-than-perfect solutions could help to establish first application paradigms which, in their turn, might help accelerate progress.

Speech translation

Despite the embattled 40-year history of language (text) translation systems, speech translation is now being researched across the world because of the realisation that far-less-than-perfect paragraph-by-paragraph translation could yield useful applications in the shorter term. In their turn, those first application paradigms could serve as drivers of further progress. The German Verbmobil project (http://verbmobil.dfki.de/), for instance, demonstrated just how difficult human-human spoken dialogue translation is. Once application paradigms have been achieved, however, speech translation technology would appear set to gain an enormous market. Still, it may take quite some time before there is a massive growth in the market for speech translation products, due to the difficulty of the research which remains to be done.

Conversational spoken dialogue

For some time, the term ‘conversational spoken dialogue’ has been a catch-all for next-step spoken language dialogue systems, such as those explored in the DARPA Communicator

project. However, the DARPA Communicator agenda remains focused on task-oriented dialogue, such as flight ticket reservation. Even if conducted through mixed initiative spoken dialogue in which the human and the machine exchange dialogue initiative in the course of their dialogue about the task, task-oriented spoken dialogue might not qualify as conversational spoken dialogue. Conversational spoken dialogue is mixed-initiative, to be sure, but in conversational spoken dialogue there is no single task and no limited number of distinct tasks which have to be accomplished. Rather, spoken conversation systems may be characterised as *topic-oriented.* It is the breadth and complexity of the topic(s) on which the system is able to conduct conversation which determine its strength. Research on spoken conversation systems is still limited. Obviously, however, spoken conversation systems hold an enormous application potential because they represent the ultimate generalisation of the qualities which everybody seem to appreciate in task-oriented mixed initiative spoken language dialogue systems.

Key technologies: multimodal systems

In addition to speech-only technologies, the timelines in Section 3 highlight a series of multimodal speech systems technologies which are still at the research stage in most cases, including:

- intelligent multimodal information presentation including speech;
- natural interactivity;
- immersive virtual reality and augmented reality.

Intelligent multimodal information presentation including speech

Intelligent multimodal information presentation including speech is a mixed bag of complex technologies which do not seem to have any clear research direction at the present time. The reason is that the term *multimodality,* as pointed out in Section 2 above, refers to a virtually unlimited space of combinations of (unimodal) modalities. Thus, Modality Theory (Bernsen 1997b, 2001) has identified an exhaustive developers' toolbox of unimodal input/output modalities in the media of graphics (or vision), acoustics (or hearing), and haptics (or touch) consisting of more than a hundred unimodal modalities. The number of possible combinations of these unimodal input/output modalities is evidently staggering and, so far, at least, no way has been found to systematically generate a subset of good and useful modality combinations which could be recommended to system developers. The best current approach is to list modality combinations which have been found useful already in experimental or development practice. Obviously, given the limited exploration of the space of possible modality combinations which has taken place so far, those combinations constitute but a tiny fraction of the modality combinations which eventually will be used in HHSI. The same lack of systematicity applies to the subset of useful modality combinations which include speech output and/or speech input. Thus, for instance, it is known that speech and static graphics image output is a useful modality combination for some purposes and that the same holds for combined speech and pen input into various output domains as well as for speech and pointing gesture input into, e.g., a static graphics map output domain. The qualifying term *intelligent* is being used to distinguish intelligent multimodal information presentation systems from traditional multimedia presentations. In traditional multimedia presentations, the user uses keyboard and mouse (or similar devices) to navigate among a fixed set of output options all of which have been incorporated into the system at design-time. In intelligent multimodal information presentation systems, the system itself generates intelligent multimodal output at run-time. This may happen through run-time language and/or speech generation coordinated with run-time graphics image generation and in many other ways as

well. Some years ago, a reference model for intelligent multimodal information presentation systems was proposed by an international consortium of developers (Computer Standards and Interfaces 18, 6-7, 1997). Since then, little systematic development has happened, it appears, which is probably due to the fact that the field is as open-ended at it is. Still, it would appear that (i) the field of intelligent multimodal information presentation systems is an extremely promising approach to complex interactive information presentation, such as in interactive systems for instruction tasks for which several output modalities are needed, including speech. In order to advance research in this field, research is needed on Modality Theory in order to identify potentially useful modality combinations as well as on next-step architectures and platforms for intelligent multimodal information presentation.

Natural interactivity

As argued in Section 2, fully natural interactive systems represent a necessary vision for a large part of the field of interactive systems. Furthermore, spontaneous speech input/output is fundamental to natural interactive systems. Given this (latter) fact, it would seem that speech research is set to take the leading role in the development of increasingly natural interactive systems. Already today, this research and development process can be broken down into a comprehensive, semi-ordered agenda of research steps. The steps include, at least, (i) *fundamental research on human communicative behaviour,* including identification of the relevant phenomena which are being coordinated in human behaviour across abstraction levels and modalities, such as speech prosody and facial expression; validated coding schemes for these phenomena; and standard tools for coding the phenomena in order to create research and training resources in an efficient and re-usable fashion; (ii) *speech and graphics integration* in order to achieve full run-time coordination of spoken output with lip movement, facial expression, gaze, gesture and hand manipulation, and bodily posture; (iii) *speech and machine vision integration* in order to enable the system to carry out run-time understanding of spoken input in combination with lip movement, facial expression, gaze, gesture and hand manipulation, and bodily posture; and (iv) *conversational spoken dialogue* as discussed above. Other relevant technologies include, i.a., machine learning and 3D graphics modelling of human behaviour. Although research in underway on (i) through (iv), there is no doubt that the field might benefit strongly from a focused effort which could connect the disparate research communities involved and set a stepwise agenda for achieving rapid progress. The application prospects are virtually unlimited, as witnessed by the consensus in the field that increased natural interaction tends to generate increased trust in HHSI.

Immersive virtual reality and augmented reality

It is perhaps less clear what are the speech technology application prospects of immersive virtual reality. Today, immersive virtual reality requires that users are wired up with 3D goggles, force feedback data gloves, data suits, and/or wired surfaces and other wired equipment, such as flight cockpits or bicycles. At the present time, it seems uncertain to which extent and for which purposes immersive virtual reality technologies will be found useful in the future. The primary purposes for which these technologies are being used to day are advanced technology exhibition and demonstration, and the building of rather expensive simulation setups, such as flight simulators. Furthermore, it is far from clear which role(s) speech will come to play in immersive virtual environments. These remarks also apply to *augmented reality* technology.

Other research and supporting measures needed

In order to promote efficient research progress on advanced interactive systems which include speech as a modality, technology research is far from sufficient. As pointed out in Section 2,

present and future advanced systems research takes place in an extremely complex context in which leading research efforts must incorporate global state-of-the-art developments in many different fields. World-leading speech-related systems research should be accompanied by the following kinds of research, at least:

- state-of-the-art generic platforms;
- generic architectures;
- hardware;
- specialised best practice in development and evaluation;
- standard re-usable resources;
- behavioural research;
- neural basis for human natural communicative behaviour;
- design of form and contents;
- porting technologies to languages, cultures and the web;
- the disabled;
- maintenance for uptake.

State-of-the-art generic platforms

In order to effectively aim at exploitable results from early on, speech-related systems research needs to build upon existing state-of-the-art generic platforms including APIs. If a state-of-the-art generic platform is not available to the researchers, either because it does not yet exist or because it is inaccessible for proprietary reasons, researchers have to build it themselves. This is not possible in small-scale research projects which have an additional research agenda which presupposes a working platform. The consequence is that the research project will either build upon some sub-optimal platform in order to complete the research agenda, or build a better platform but not complete the research agenda. Both consequences are unacceptable, of course, but the former may work temporarily if the research aims are very advanced ones. However, when the research aims have been achieved or, at least, somehow explored, there will typically be no practical way of continuing the research in order to produce a state-of-the-art generic platform which could bring the research results towards the market. Two implications seem to follow: (i) it would be highly desirable if companies could be encouraged to make their most advanced platforms accessible to researchers. (ii) If a state-of-the-art generic platform is missing altogether, it should either be produced in a separate project or projects should be made so large as to include platform development. Both implications would seem to require a transformation of existing European research funding mechanisms.

Generic architectures

It would seem likely that overall research speed and efficiency in Europe could be accelerated by research on *generic architectures* for future systems, such as conversational spoken dialogue systems, intelligent multimodal information presentation systems which include speech, or natural interactive systems. In the absence of research initiatives on generic architectures for future systems, research projects are likely to specify idiosyncratic architectures which may satisfy their present needs but which do not sufficiently take into account global developments nor prepare for the next steps in advanced systems development. For the time being, there does not appear to be any European speech-related initiative in this field apart from the CLASS project which was launched in the autumn of 2000 (http://www.class-tech.org/). For efficiency, work on generic architectures should be done as

a collaborative effort between many small-scale research projects and industry as in CLASS, or between a medium-scale research project and industry.

Hardware

Increasingly, advanced systems demonstrators require *hardware* design and development. For many research laboratories, this is a new challenge which they are ill-prepared to meet. Moreover, there is no strong tradition for involving hardware producers in the field of speech technologies, primarily because the need for involving them is a rather recent one. Ways must be found to forge links with leading hardware producers in order to make emerging hardware available to researchers. This problem has much in common with the platform issue discussed above.

Specialised best practice in development and evaluation

Advanced speech systems research is conducted in a software engineering space bounded by, on the one hand, general software engineering best development and evaluation practice and, on the other, emerging ISO standards and de facto standards imposed by global industrial competition. Between these boundaries lies software engineering best practice in development and evaluation specialised for various speech-related systems and component technologies. This field remains ill-described in the literature. Apart from the DISC project on best practice in the development and evaluation of spoken language dialogue systems (www.disc2.dk), some work on evaluation in EAGLES Working Groups during the 1990s (http://www.ilc.pi.cnr.it/ EAGLES96/home.html), various national evaluation campaigns, and planned work in CLASS, little work has been done in Europe. By contrast, massive work has been done on component evaluation in the US over the last fifteen years. The result is that the speech-related technology field is replete with trial and error, repetitions of mistakes, and generally sub-state-of-the-art approaches. These negative effects are multiplied by the presence in the field of a large number of developers who are new to the field.

Admittedly, the field of software engineering best practice in development and evaluation specialised for various speech systems and component technologies is difficult and costly to do something about under present conditions. Technology *evaluation* campaigns are costly to do and require serious logistics. Yet the US experience would seem to indicate that technology evaluation campaigns are worth the effort if carried out for key emerging technologies including some of the technologies described in this paper. When a technology has gone to the market, industry does not want to participate any more and rather wants, e.g., evaluation toolkits for internal use. For emerging technologies, however, technology evaluation campaigns are an efficient means of producing focused progress. In fact, all participants tend to become winners in the campaigns irrespective of their comparative scorings according to the metrics employed, because everybody involved learns how to improve, or when to discard, their technologies and approaches. For Europe, technology evaluation campaigns for key emerging technologies could be a means of creating lasting advances on its global competitors. In order to take care of the complex logistics needed for the campaigns, it is worth considering to establish a European agency similar to the US NIST (National Institute for Standards in Technology) whose comprehensive experience with technology evaluation campaigns makes it comparatively easy to plan and launch campaigns in novel emerging technologies. Alternatively, NIST might be asked to undertake to run technology development and evaluation campaigns in Europe, provided that this does not offend political and industrial sensibilities too much.

Effective *development* best practice work specialised for speech technologies is difficult to do under the current European funding mechanisms. The reason is that development best practice work requires access to many different components, systems and approaches in order to

create an effective environment for the discussion and identification of best practice. This environment can only be established across many different small-scale projects or within medium-scale projects. CLASS is the first example of such an environment.

Standard re-usable resources

The term *resources* covers raw data resources, annotated data resources, annotation schemes for data annotation, and annotation tools for efficient automatic, semi-automatic or manual annotation of data. Resources are crucial for many different purposes, such as research into coding schemes or the training of components. Also, resources tend to be costly to produce. This means that, if the relevant resources are not available, research projects often take the easy way out which is to use less relevant but existing and accessible resources for their research. The results are sub-optimal research results and slowed-down progress. Common to resources of any kind is the need for standardisation. If some resource is not up to the required standards, its production is often a waste of effort because the created resource cannot be used for anything useful. In its strategy paper from 1991, ELSNET (http://www.elsnet.org/) proposed the establishment of a European resources agency. This recommendation was adopted through the creation of ELRA (European Language Resources Agency http://www.icp.inpg.fr/ELRA/ home.html) in 1995. ELRA is now a world-recognised counterpart to the US LDC (Linguistic Data Consortium, http://www.ldc.upenn.edu/). Still, ELRA is far from having the capacity to produce on its own all the resources and standards needed for efficient research progress. By contrast with technology evaluation campaigns, Europe has been active in the resources area during the 1990s. Today, there is a strong need to continue activities in producing publicly available resources and standards for advanced natural language processing, natural interactive systems development, evaluation campaigns as described above, etc. Recently, the ISLE (International Standards for Language Engineering) Working Group on Natural Interactivity and Multimodality (http://www.isle.nis.sdu.dk) has launched cross-Atlantic collaboration in the field of resources for natural interactivity and multimodality.

Behavioural research

Humans are still far superior to current systems in all aspects of natural interactive communication. Furthermore, far too little is known about the natural interactive behaviour which future systems need to be able to reproduce as output or understand as input. There is a strong need for basic research into human natural communicative behaviour in order to chart the phenomena which future systems need to reproduce or understand. This research will immediately feed into the production of natural interactive resources for future systems and components development, as described above.

Neural basis for human natural communicative behaviour

Related to, but distinct from, basic research into human natural communicative behaviour is basic research into the neural basis for human natural communicative behaviour. In the heydays of cognitive science in the 1980s, many researchers anticipated steady progress in the collaboration between research on speech and language processing, on the one hand, and research into the neural machinery which produces human speech and language on the other. However, massive difficulties of access to how human natural communicative behaviour is being produced by the brain turned out to prevent rapid progress in linking neuroscience with speech and language processing research. Today, however, due to the availability of technologies such as MR imaging and PET scanning, as well as the increasing sophistication of the research agenda for the speech technology field, the question arises if it might be timely to re-open the cognitive science agenda just described. Potential results include, among others, input to generic architecture development (cf. above), identification of biologically

motivated units of processing, such as speech and lip movement coordination, and identification of biologically motivated modalities for information representation and exchange. Relevant research is already going on in the field of neuroscience but, so far, few links have been established to the fields of speech technologies and natural interactive systems more generally.

Design of form and contents

Yet another consequence of the increasing emphasis on systems as opposed to system components is the growing importance of form and contents design. It is a well-established fact that design and development for the web requires skills in contents design and contents expression which are significantly different from those which have been developed through centuries for text on paper. In order to develop good demonstrator systems for the web or otherwise, there is a need for strongly upgraded skills in the design and expression of multimodal digital contents. For instance, it is far from sufficient to have somehow gleaned that speech might be an appropriate modality for some intelligent multimodal information presentation instruction system and to have available a state-of-the-art development platform for building the system. To actually develop the system, professional expertise in form and contents design is required. At the present time, few groups or projects in the speech field are adequately staffed to meet this challenge.

Porting technologies to languages, cultures and the web

Right now, the gap between the "have" countries whose researchers have access to advanced speech and natural interactivity components and platforms, and the "have-not" countries whose researchers cannot use those technologies for their own purposes because they speak different languages and behave differently in natural interactive communication, seems to be increasing. There is therefore a need to *port advanced technologies to different languages and cultures* both in Europe and across the world. The market will close the gap eventually in its own way, of course. However, in order to rally the full European research potential in the field in a timely fashion, it would appear necessary to actively stimulate the porting of technologies to new languages and cultures. From a research point of view, the best way to make this happen might be to include in medium-to-large-scale projects the best researchers from "have-not" countries even if, by definition, those researchers have to spend significant time catching up on basic technologies and resources before being able to actively contributing to the research agenda.

There is another sense of the 'porting technologies' -phrase in which Europe as a whole risks falling behind global developments. It is that of *porting speech, multimodal and natural interactivity technologies to the web.* The claim here is not that this is not happening already. The claim is that this cannot happen fast enough. In order to increase the speed of porting technology to the web, it would seem necessary to strongly promote advanced components and systems development for the web. It is far from sufficient to wait until some non-speech technology has been marketed for the web, such as electronic commerce applications, and then try to "add speech" to the technology. A much more pro-active stance would appear advisable, including a strongly increased emphasis on form and contents design as argued above.

The disabled

Advanced technologies for the disabled have a tendency to lag behind technology development more generally for the simple reason that the potential markets for technologies for the disabled are less profitable. Correspondingly, advanced technologies development for the disabled tends to be supported by small separate funding programmes rather than being integrated into mainstream programme research. In many cases, however, it would appear that

systems and components technologies could be developed for any particular group of users before being transferred into applications for many other user groups. To the extent that this is the case, there may be less of a reason to confine the development of technologies for the disabled to any particular research sub-programme.

Maintenance for uptake

Finally, the small-scale science paradigm of small and isolated research projects does not at all cater for the fact that, in the complex world of advanced systems research, a wealth of prototype systems, proto-standard resources, web-based specialised best practice guides, etc., are being produced which have nowhere to go at the end of the projects in which they were developed. Their chances of industrial uptake, re-use by industry and research, impact on their intended users, etc., might become very substantially increased if it were possible to maintain them and make them publicly accessible for, say, two years after the end of projects. For this to happen, there is a need for (i) a stable web portal which can host the results, such as the present HLT (Human Language Technologies) portal under development (http://www.HLTCentral.org); (ii) open source clauses in research contracts for technologies which have nowhere to go at the end of a project; and (iii) financial support for maintenance. These requirement are likely to impose considerable strain of current European research support mechanisms. However, with some legal effort and a modest amount of financial support, the many research results produced in the speech-related field in Europe which are not being taken up immediately and which are not within the remit of ELRA, could gain much more impact than is presently the case.

5. Proposed Actions

Early preparations for the European Commission's 6th Framework Programme (FP6) including IST (Information Society Technologies) research are now in progress. It is premature to make predictions with any degree of certainty as to how the IST part of FP6 will shape up. Current information suggests an increased emphasis on basic research compared to the present FP5. In addition, it is possible that FP6 will include opportunities for the medium-scale research initiatives which were called for on several occasions above, i.e. large-scale "clusters" of projects all addressing the same research topic in a coordinated fashion. Finally, the current covering title for FP6 IST research is "ambient intelligence" which is one of the terms of fashion quoted in the present paper. Given the timelines and their analysis above, it does not seem to matter much which covering term is being chosen for FP6. "Ambient intelligence" is as apt as several others for FP6 and future advanced interactive systems research but, as argued in Section 3, it is far from clear if ambient intelligence requires us to focus on any particular segment of future speech-related technologies. However, the possible, increased emphasis on basic research as well as the possibility of carrying out medium-scale science in speech-related technologies are to be welcomed in the light of the argument above.

5.1 Research priorities for speech-related technologies 2000-2010

Taking into our stride the transformations of the field of speech-related research from speech-only to interactive systems in general, and from components research to interactive systems research, the top priorities in speech-related technologies research are:

- multi-speaker meeting transcription development and evaluation campaigns;
- speech summarisation development and evaluation campaigns;
- speech translation prototypes, generic platforms, and generic architectures. Development and evaluation campaigns are highly desirable;

- conversational spoken dialogue prototypes, generic platforms, and generic architectures. Development and evaluation campaigns are highly desirable;
- next-step prototypes, generic platforms, and generic architectures for intelligent multimodal information presentation;
- next-step prototypes, generic platforms, and generic architectures for natural interactive systems.

As soon as theoretically and practically feasible, all of the above advanced speech, multimodal and natural interactivity technologies should be developed for the web including hardware, form and contents design. The fact that some top research priorities have been mentioned above emphatically does not preclude the desirability of continuing "business as usual" in the field of speech-related research, including continued research into *all* of the technologies which have been mentioned earlier in the present paper. On the contrary, business as usual is actually assumed by the above top priorities list which focuses on technologies over and above business as usual. This also applies to next-step research into already deployed speech-related technologies, such as mixed initiative, task-oriented spoken dialogue systems.

For basic research leading to novel concepts, theories and formalisations, the top priorities are:

- basic research into human natural communicative behaviour;
- a novel theory of natural communication which can replace speech acts theory and discourse theory by taking the notion of a complete communicative act as its basic notion;
- research on Modality Theory in order to identify potentially useful modality combinations;
- establishment of collaborative links to research into the neural basis for human natural communicative behaviour.

5.2 Research organisation needed

Medium-scale science is needed for, at least, the coordinated development of natural interactive systems prototypes, generic platforms, generic architectures, best practice in development and evaluation, and standard resources. A large, medium-scale science project with these objectives should include the porting of technologies to new languages and cultures.

It is quite possible that the medium-scale science model could be applied to research into other speech-related technologies, such as speech translation technologies, conversational spoken dialogue systems, or speech technologies for ambient intelligence.

For researchers in small-scale speech-related projects, in particular, the creation of a generic platforms and hardware "bourse" through contributions from European industry would be of great importance.

Finally, we should stop having research programme ghettos for technologies for the disabled.

5.3 Infrastructural actions needed

In order to promote maximum uptake of the research results produced, it would be highly desirable to have funding for low-cost ways of maintaining research results for later uptake.

Given the emphasis on technology development and evaluation campaigns above, Europe needs to establish an evaluation and standards agency. It is not evident to the present author

that current political and industrial sensibilities would allow the US NIST to undertake to run technology development and evaluation campaigns in Europe.

This having been said, there is much to be said for increasing global collaboration on many aspects of speech-related research, such as creating a coordinated global infrastructure for resources distribution.

References

Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., and Suhm, B.: Audio-Visual and Multimodal Speech-Based Systems. In D. Gibbon, I. Mertens and R. Moore (Eds.): *Handbook of Multimodal and Spoken Dialogue Systems*. Dordrecht: Kluwer Academic Publishers 2000, 102-203.

Bernsen, N. O.(1997a): Towards a tool for predicting speech functionality. *Speech Communication* 23, 1997, 181-210.

Bernsen, N. O. (1997b): Defining a Taxonomy of Output Modalities from an HCI Perspective. *Computer Standards and Interfaces,* Special Double Issue, 18, 6-7, 1997, 537-553.

Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. In Granström, B. (Ed.): *Multimodality in Language and Speech Systems.* Dordrecht: Kluwer Academic Publishers 2001 (to appear).

CLASS: http://www.class-tech.org/

Computer Standards and Interfaces, Special Double Issue, 18, 6-7, 1997.

DARPA Communicator: http://fofoca.mitre.org/

DISC www.disc2.dk

EAGLES: http://www.ilc.pi.cnr.it/EAGLES96/home.html

ELRA: http://www.icp.inpg.fr/ELRA/home.html

ELSNET http://www.elsnet.org/

i3: http://www.i3net.org/

ISLE: http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

ISLE Working Group on Natural Interactivity and Multimodality: http://www.isle.nis.sdu.dk

HLT portal: http://www.HLTCentral.org

LDC http://www.ldc.upenn.edu/

SmartKom: http://smartkom.dfki.de/start.html

Verbmobil: http://verbmobil.dfki.de/

Towards a Road Map on Human Language Technology: Natural Language Processing

Editors: Andreas Eisele, Dorothea Ziegler-Eisele

Version 2 (March 2002)

Abstract

This document summarizes contributions and discussions from two workshops that took place in November 2000 and July 2001. It presents some visions of NLP-related applications that may become reality within ten years from now. It investigates the technological requirements that must be met in order to make these visions realistic and sketches milestones that may help to measure our progress towards these goals.

1. Introduction

Scope of this Document

One of the items on ELSNET's agenda for the period 2000-2002 is to develop views on and visions of the longer-term future of the field of language and speech technologies and neighboring areas, also called ELSNET's Road Map for Human Language Technologies. As a first step in this process, ELSNET's Research Task group is organizing a series of brainstorming workshop with a number of prominent researchers and developers from our community. The first one of these workshops took place in November 2000 under the general motto “How will language and speech technology be used in the information world of 2010? Research challenges and infrastructure needs for the next ten years". The second one was co-organized in July 2001 by ELSNET and MITRE as part of ACL-2001 and had the somewhat more specific orientation on “Human Language Technology and Knowledge Management (HLT-KM)”. This workshop brought together more than 40 researchers from industry and academia and covered a considerable range of topics related to KM and HLT in general.

This paper aims at summarizing and organizing material from both workshops, but concentrates on applications and technologies that involve NLP, i.e. the processing of written natural language, as speech-related technologies and new models of interactivity have already been covered in documents presented around the first workshop. In the discussion of question answering and summarization, vision papers and roadmaps compiled by researchers in the US and published by NIST have been taken as an additional source of inspiration.

The Growing Need for Human Language Technology

Natural language is the prime vehicle in which information is encoded, by which it is accessed and through which it is disseminated. With the explosion in the quantity of on-line

text and multimedia information in recent years there is a pressing demand for technologies that facilitate the access to and exploitation of the knowledge contained in these documents.

Advances in human language technology will offer nearly universal access to on-line information and services for more and more people, with or without skills to use computers. These technologies will play a key role in the age of information and are cited as key capabilities for competitive advantage in global enterprises.

Extraction of knowledge from multiple sources and languages (books, periodicals, newscasts, satellite images, etc.) and the fusion into a single, coherent textual representation requires not only an understanding of the informational content of each of these documents, the removal of redundancies and resolution of contradictions. Also, models of the user are required, the prior knowledge that can be assumed, the level of abstraction and the style that is appropriate to produce output that is suitable for a given purpose.

More advanced knowledge management (KM) applications will be able to draw inferences and to present the conclusions to the user in condensed form, but let the user ask for explanations of the internal reasoning. In order to find solutions for problems beyond a static pool of knowledge, we need systems that are able to identify experts, who have solved similar problems. Again, advanced NLP capabilities will be required to appraise the aptitude of candidates from documents authored by them or describing prior performance.

But also outside of KM, sophisticated applications of NLP will emerge over the next years and decades and find their way into our daily lives. The range of possibilities is almost unlimited. An important group of applications is related to electronic commerce, i.e. new methods to establish and maintain contact between companies and their customers. Via mobile phones, e-mail, animated web-based interfaces, or innovative multi-channel interfaces, people will want to make use of all kinds of services related to buying and selling goods, home-banking, booking of journeys, and the like. Also in the area of electronic learning a considerable growth is expected within the coming years.

Multilinguality

Whereas English is still the predominant language on the WWW, the fraction of non-English Web pages and sites is steadily increasing. Contrasting earlier apprehensions, the future will probably present ample opportunities for giving value to different languages and cultures. However, the possibility to collect information from disparate, multilingual sources also provides considerable challenges for the human user of these sources and for any kind of NLP technology that will be employed.

One of the major challenges is lexical complexity. There will be about 200 different languages on the web and thus about 40.000 potential language pairs for translation. Clearly, it will not be possible to build bilingual dictionaries that are comprehensive both in the number of language pairs and in the coverage of application domains. Instead, multilingual vocabularies need to provide mappings into language independent knowledge organization structures, i.e. common systems of concepts linked by semantic relations. However, the definition of such an “interlingua” will be difficult in cases in which languages make distinctions of different granularity.

Research Trends and Challenges

The field of human language technology covers a broad range of activities with the goal of enabling people to communicate with machines using natural communication skills.

Although NLP can help to facilitate knowledge management, it requires a large amount of specialized knowledge by itself. This knowledge may be encoded in complex systems of linguistic rules and descriptions, such as grammars and lexicons, which are written in dedicated grammar formalisms and typically require many person-years of development effort. The rules and entries in such descriptions interact in complex ways, and adaptation of such a sophisticated system to a new text style or application domain is a task that requires a considerable amount of specialized manpower.

One way to cope with the difficulties in the acquisition of linguistic knowledge was to restrict attention to shallower tasks, such as looking for syntactic "chunks" instead of a full syntactic analysis. Whereas this has proven rather successful for some applications, it obviously severely limits the depth to which the meaning of a document or utterance is taken into account.

Another approach was to shift attention towards models of linguistic performance (what occurs in practice, instead of what is principally possible) and to use statistical or machine learning methods to acquire the necessary parameters from corpora of annotated examples. These data-driven approaches offer the possibility to express and exploit gradual distinctions, which is quite important in practice. They are not only easier to scale and adapt to new domains, their algorithms are also inherently robust, i.e. they can deal, to a certain extent, gracefully with errors in the input.

Statistical parsers, trained on suitable tree banks, now achieve more than 90% precision and recall in the recognition of syntactic constituents in unseen sentences from English financial newspaper text.

However, a lot of work remains to be done, and it is not obvious how the success of corpus-driven approaches can be enlarged along many dimensions simultaneously. One challenge is that analysis methods need to work for many languages, application domains and text types, whereas the manual annotation of large corpora of all relevant types will not be economically feasible. Another challenge is that, other than syntax, many additional levels of analysis will be required, such as the identification of word sense, the reference of expressions, structure of argumentation and of documents, and the pragmatic role of utterances. Often, the theoretical foundation that is required before the annotation of corpora can begin is still lacking.

One could say that for corpus-driven approaches the issue of scalability of the required resources shows up again, albeit in a somewhat different disguise. Hence, research in NLP will have to address this issue seriously, and find answers to the question how better tools and learning methods can reduce the effort of manual annotation, how annotated corpora of a slightly different type could best be re-used, how data-driven acquisition processes can exploit and extend existing lexicons and grammars, and finally how analysis levels for which the theoretical basis is still under development could be advanced in a data-driven way.

Structure of this Document

The remainder of this document is structured as follows. In Chapter 2 we describe a number of prototypical applications and scenarios in which NLP will play a crucial role. Whereas each of these scenarios is discussed mainly from a user's perspective, we also give indications, which technological requirements must be met to make various levels of sophistication of these applications possible. In Chapter 3, the technologies that have been mentioned earlier are discussed in more detail, and we try to indicate which levels of functionality may be expected within the timeframe of this study. These building blocks are

then put into a tentative chronological order, which is displayed in Chapter 4. Finally, Chapter 5 gives some general recommendations about beneficial measures concerning the infrastructure for the relevant research.

2. Applications of NLP

Recent developments in natural language processing have made it clear that formerly independent technologies can be harnessed together to an increasing degree in order to form sophisticated and powerful information delivery vehicles. Information retrieval engines, text summarizers, question answering and other dialog systems, and language translators provide complementary functionalities which can be combined to serve a variety of users, ranging from the casual user asking questions of the web to a sophisticated, professional knowledge worker.

Though one cannot strictly separate the following applications from each other, because one can act as a part of another, we try to dissect the large field of existing and future applications in the hope of making the field as a whole more transparent.

Information Retrieval (IR)

What is called information retrieval today is actually but a foretaste of what it should be. Current systems neither understand the information need of the user, nor the content of the documents in their repositories. Instead of meaningful replies, they just return a ranked, and often very long list of documents that are somehow related to the given query, which is typically very short. A better name for this restricted functionality would be text retrieval.

Information retrieval systems must understand a query, retrieve relevant information, and present the results. Retrieved information may consist of a long document, multiple documents of the same topic, etc and good systems should present the most important material in a clear and coherent manner.

Current information retrieval techniques either rely on an encoding process using a certain perspective or classification scheme to describe a given item, or perform a superficial full-text analysis, searching for user-specific words. Neither case guarantees content matching.

The ability to leverage advances in input processing (especially natural language query processing) together with advances in content-based access to multimedia artifacts (e.g., text, audio, imagery, video) promises to enhance the richness and breadth of accessible material while at the same time improving retrieval precision and recall and thus reducing the search time. Dealing with noisy, large scale, and multimedia data from sources as diverse as radio, television, documents, web pages, and human conversations (e.g., chat sessions and speech transcriptions) will offer challenges.

One important part of IR would be multi-document summarization that can turn a large set of input documents into several different short summaries, which can then be sorted by topics or otherwise put into a coherent order.

Summarization

Summarization will enable knowledge workers access to larger amounts of material with less required reading time. The goal of automatic text summarization is to take a partially structured source text, extract information content from it and present the most important content in a condensed form in a manner sensitive to the needs of the user and task. Scalability to large collections and the generation of user-tailored or purpose-tailored summaries are active areas of research.

The summarization can either be an extract consisting entirely of material copied from the input, or an abstract containing material not present in the input, such as subject categories, paraphrases of content, etc.

For extraction shallower approaches are possible, as frequently the sentences may be extracted out of context. The transformation here involves selecting salient units and synthesizing them with the necessary smoothing (adjusting references, rearranging the text…). Training by using large corpora is possible.

Abstracts need a deeper level of analysis, the synthesis involves natural language generation and some coding for a domain is required.

Depending on their function, three types of abstracts can be distinguished: An indicative abstract provides a reference function for selecting documents for more in-depth reading. An informative abstract covers all the salient information in the source at some level of detail and evaluative abstracts express the abstractor's views on the quality of the work of the author.

Characteristics for the summarization are the reduction of the information content (compression rate), the fidelity to the source, the relevance to the user's interest, and the well-formedness regarding both to syntactic and discourse level. Extracts need to avoid gaps, dangling anaphora, ravaged tables and lists, abstracts need to produce grammatical, plausible output.

Some current applications of summarization are:

1. Multimedia news summaries: watch the news and tell what happened while I was away
2. Physicians' aids: summarize and compare the recommended treatments for this patient
3. Meeting summarization: find out what happened at that teleconference I missed
4. Search engine hits: summarize the information in hit lists retrieved by search engines
5. Intelligence gathering: create a 500-word biography of Osama bin Laden
6. Hand-held devices: create a screen-sized summary of a book
7. Aids for the Handicapped: compact the text and read it out for a blind person

Though there are already promising approaches towards mastering all types of summaries, there are still obstacles to overcome such as the need for robust methods for the recognition of semantic relations, speech acts, and rhetorical structure.

Question Answering (QA)

The straightest way to get access to the gigantic volume of knowledge around us is probably asking questions by communicating with other persons, computers or machines.

An important new class of systems will move us from our current form of search on the web (type in keywords to retrieve documents) to a more direct form of asking questions in natural language, which are then directly responded to with an extracted or generated answer. Currently it is rather straightforward to get an answer to "what questions" (what is the capital of China, what are the opening hours of the hermitage etc.), whereas "why questions" (why did the new market fail) are normally not answered by an information retrieval query, unless the answer happens to be present in the information database, or can be inferred afterwards by the user from the answers she gets.

In the next decade time has come to find answers to why questions from information systems by letting the systems make the appropriate inferences. This requires very sophisticated automatic reasoning methods, based on systematic extraction of information from texts, storing the information in a systematized way, which lends itself to reasoning and inference rules that will be able to draw the proper conclusions from the knowledge stored in the information database.

We can subdivide the long-term goal of building powerful, multipurpose information management systems for QA in simpler subtasks that can be attacked in parallel at varying levels of sophistication, over shorter time frames.

Clearly there is not a single, archetypical user of a Q&A system. In fact there is a full spectrum of questions, starting with simple factual questions, which could be answered in a single short phrase found in a single document (e.g. "Where is the Taj Mahal?"). Next, questions like "What do we know about Company xyz?", where the answer cannot be found in a single document but will require retrieving multiple documents, locating portions of answers in them and combining them into a single response. This kind of question might be addressed by decomposing it into a series of single focus questions.

Finally there are very complex questions, with broad scope, using judgment terms and needing deep knowledge of the user's context to be answered. Imagine someone is watching a television newscast, becomes interested in a person, who appears to be acting as an advisor to the country's Prime Minister. And now the person wants to know things like: "Who is this individual. What is his background? What do we know about the political relationship of this person and the Prime Minister and/or the ruling party?". The future systems that can deal with this type of questions must manage the search in multiple sources in multiple media/languages, the fusion of information, resolution of conflicting data, multiple alternatives, adding interpretation, drawing conclusions.

In order to realize this goal, research must deal with question analysis, response discovery and generation from heterogeneous sources, which may include structured and unstructured language data of all media types, multiple languages, multiple styles, formats and also image data i.e. document images, photography and video.

To the extent to which NLP research will learn to master the challenges of source selection, source segmentation, extraction, and semantic integration across heterogeneous sources of unstructured and semi-structured data, NLP technology will help us to reduce the time,

memory, and attention required to sift through many returned web pages from a traditional search by providing direct answers to questions.

Semantic Web

The standardization committee for the WWW (called W3C) expects around a billion web users by 2002 and an even higher number of available documents. However, this success and exponential grow makes it increasingly difficult to find, to access, to present, and to maintain the information of use to a wide variety of users.

The semantic web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.

The semantic web is not a separate web but an extension of the current one, in which information is given well-defined meaning better enabling computers and people to work in cooperation. With the help of ontologies large amounts of text can be semantically annotated and classified.

Currently pages on the web use representations rooted in format languages such as HTML or SGML. The information content, however, is mainly presented by natural language. Thus, there is a wide gap between the information available for tools that try to address the problems above and the information kept in human readable form.

The semantic web will provide intelligent access to heterogeneous and distributed information enabling software agents to mediate between the user needs and the available information sources.

The first steps in weaving the semantic web into the structure of the existing web are already under way. In the near future, these developments will usher in significant new functionality as machines become much better able to process and "understand" the data that they merely display at present.

What is required: creation of a machine understandable semantics for some or all of the information presented in the WWW i.e.

- Developing languages for expressing machine understandable meta-information for documents, in the line of RDF, DAML, and similar proposals.
- Developing terminologies (i.e., name spaces or ontologies) using these languages and making them available on the web.
- Integrating and translating different terminologies
- Developing tools that use such languages and terminologies to provide support in finding, accessing, presenting and maintaining information sources.

Developing such languages, ontologies and tools is a wide-ranging problem that touches on the research areas of a broad variety of research communities.

Creation of the relevant tools will require a better knowledge of what the users want to know from websites, i.e. these developments need to be based on a user-centered process view.

Another crucial issue will be: “Who is going to populate the semantic web?” The semantic markup that is required by automated software agents needs to be very easy to create and supporting tools need to be provided, otherwise this wonderful idea will not have significant impact for a long time. Advanced NLP technology that can “guess” the correct semantic annotation and propose suitable markup semi-automatically will enable conformance to the needs of software agents with minimal manual effort.

Dialogue Systems

No matter if people want to buy something, find or use a service or just need information, dialog systems promise user-friendly and effective ways to achieve these goals, even for first time users.

Despite the apparent resemblance to QA systems, there are several specific problems to be solved concerning dialogue modality and structure. Input to a dialog system might be via keypad, voice, pointing device, combinations thereof, or other channels, so all errors and incompleteness of spontaneous natural language will show up. In contrast to QA systems, there will be mixed initiatives of speaker and system and the scope is much wider if we take into account that the focus during natural dialogue may often change. Also, the utterance made during a dialog can only be correctly interpreted based on the dialog context and the mutual knowledge that has been accumulated before it was made.

In future we require systems that can support natural, mixed initiative human computer interaction that deals robustly with context shift, interruptions, feedback and shift of locus or control.

Open research challenges include the ability to tailor flow and control of interactions and facilitate interactions including error detection and correction tailored to individual physical, perceptual and cognitive differences.

Motivational and engaging life-like agents offer promising opportunities for innovation.

Agent/user modeling: Computers can construct models of user beliefs, goals and plans as well as models of users’ individual and collective skills by processing materials such as documents or user interactions/conversations. While raising important privacy issues, modeling users or groups of users unobtrusively from public materials or conversations can enable a range of important knowledge management capabilities

tracking of user characteristic skills and goals enhances interaction as well as discovery of experts by other users or agents

A central problem for the development of dialogue systems is the fact that contemporary linguistics is still struggling to achieve a genuine integration of semantics and pragmatics. A satisfactory analysis of dialogue requires in general both semantic representation i.e. representation of the content of what the different participants are saying and pragmatic information, i.e. what kinds of speech acts they are performing (are they asking a question, making a proposal…)

Analysis of a dialog needs to explain the purpose behind the utterances it consists of. Determining the semantic representation of an utterance and its pragmatic features must in general proceed in tandem. A dialogue system identifying the relevant semantic and pragmatic information will thus have to be based on a theory in which semantics and

pragmatics are both developed with the formal precision that is a prerequisite for implementation and suitably attuned to each other and intertwined.

Applications in Electronic Commerce

New technological possibilities can quickly impact the interaction between companies and their customers. One example are dialog systems that allow customers to obtain personal advises or services. For reasons indicated above, these systems are difficult to build, but once this investment has been done, they can be operated at low cost for the company.

Another example, which may be even sooner to come, is the creation of systems that support processing of emails sent by customers. According to business analyses, e-mail has already now become one of the most common forms of customer communication. For numerous businesses that are not well-prepared, this has transformed e-mail into a severe pain point, giving rise to the pressing need to adopt e-mail response management systems.

Obviously, NLP technologies that are able to extract the salient facts from email messages can constitute a central part of these systems. Due to the potential complexity of the queries and additional problems like ungrammatical input and spelling errors, the correct interpretation of arbitrary messages is far from easy. However, there are several factors that alleviate the situation: Messages that are too difficult for automatic processing can be routed to human agents. In cases in which doubts about the correctness of generated responses persist, these responses can always be checked by manual inspection. Historical data about email exchange with customers can be used to bootstrap the models that are required for the system. Depending on the business, a significant fraction of the emails may be amenable to NLP, including requests for information material, business reports, certificates, statements of account, scheduling requests, conference registrations etc.

e-Learning

Using modern technology to facilitate learning is one of the most promising application domains of NLP. Good QA systems that are able to give answers to the point, or summarization systems that can adapt to the user's prior knowledge and present important additions in a way that is easy to understand could immediately take the place of a good teacher, which an unlimited supply of time and patience. One technology is ripe to build these tools, using them for e-learning will one of the biggest opportunities to our knowledge society.

However, as the European society evolves more and more into multilingualism, it is natural to ask how NLP can help to make language learning easier and more effective. We can imagine systems to help train children to write and to speak a foreign language. There will be combinations of multi-modal aids for the handicapped. A child will write a sentence and the system will correct it and tutor him about the problems. A child will read a text aloud and the system will monitor which words are not right and why and will analyze where the pronunciation problems are. Later the system would suggest some pronunciation exercises in the particular problem.

Systems that are able to guess the intention of a speaker from the speaker's utterances in a flexible and intelligent way will offer a plethora of possibilities for e-learning. As similar capabilities are required for dialog systems in general, there will be significant synergy effects between these fields of research.

Translation

The idea of machine translation (MT) has been one of the driving forces in the early days of NLP. However, even after more than 50 years of effort, current systems still produce output of limited quality, which is suitable for assimilation of foreign-language documents, but not for the production of publishable material. But even if the old dreams did not come true, MT will play an increasing role in the multilingual world.

Last year, for the first time, English constituted less than half the material on the web. Some predict that Chinese will be the primary language of the web by 2007. Given that information on the web will increasingly appear in foreign languages and not all users will be fluent in those languages, there will be a need to gist or skim content for relevance assessment and/or provide high quality translation for deeper understanding. Some forms of translation for information access is already today available in the web at no cost. The increasing demand for these services will give a push to improve their quality and the providers will find ways to increase vocabularies and translation quality semi-automatically from terminological resources, bilingual corpora and similar sources. Also the need for interactive systems that can give rough translations of chat sessions in real time will create interesting challenges.

Clearly, any systematic collection of lexical and terminological information in the form of domain-specific ontologies will help to build better MT systems for these domains. Conversely, the construction of ontologies can be facilitated by automatic alignment of existing translations, as this will naturally lead to a clustering of the vocabulary along the relevant semantic distinctions.

These developments will also have an impact on improved systems for high-quality translation for the dissemination of documents. Chances are that hybrid combinations of symbolic and stochastic translation engines, able to learn relevant terminology from translation memories will eventually achieve a level of performance that will make them useful for the professional translator. Combined with multi-modal workbenches where voice input, keyboard and mouse interaction will make the composition of the target text as convenient as possible, these new technologies may help at least in some easier domains, where so far the effort of the human translator is dominated by low-level activities such as entering the text, adjusting the formatting, copying names and numbers, which are clearly amenable to partial automation.

3. Technologies for NLP

This chapter contains a more detailed discussion of some of the technologies that are required for the applications mentioned in the last chapter. Most of the material is organized along traditional fields of research in NLP, describing technologies that already exist, but must be further developed to achieve the ambitious goals. Some technologies cannot be assigned to one specific level, because they serve a more generic purpose, such as the extraction of relevant knowledge from text corpora.

Low-Level Processing

Most systems that analyse natural language text typically start by segmenting the text into meaningful tokens. Sometimes, the exact spelling of these tokens needs to be brought into a

canonical form, so that it can match with a lexical entry. Both processes can be based on matching the input against regular expressions, for which efficient algorithms exist. Whereas this task looks straightforward from the distance, there are actually some subtle details that need to be considered. Quite often, a decision whether a word should be split at a special character or whether a dot ends a sentence or is part of the preceding word depends on the vocabulary of the domain and on layout conventions used in this document, so that general rules cannot be defined. Documents that need to be analyzed may contain markup from text processors, which needs to be stripped or interpreted in a suitable way. The knowledge required in these preliminary stages of processing can already be quite specific, so that a manual creation of suitable rule systems is not economically feasible.

Current research on the automatic tokenization and normalization of texts therefore concentrates on the question how the knowledge required by these methods can automatically be derived from examples, using techniques statistical or machine learning approaches.

Another difficulty is the treatment of noise in the input. Output of speech recognition systems often contains recognition errors at rather high rates. Utterances entered interactively or printed documents that have undergone OCR have similar problems. Unfortunately, the distortion of even a single character can mess up the linguistic analysis of the complete input. But of course, we expect NLP systems to deal gracefully and intelligently with small distortions and errors in the input.

To make systems more robust against noisy input, probabilistic techniques for the restoration of distorted signals, which have shown to be quite effective in speech recognition, need to be adapted and generalized to new applications. However, training simple-minded statistical models on massive amounts of data will often not be feasible. By now, statistical language models that incorporate grammatical knowledge are able to give slight improvements over n-gram approaches, and it seems plausible to expect that future improvements of these will be easier to use in specific situation where training data is scarce. Large vocabularies, many types of distortions, and the need to use fine-grained contextual knowledge for improved predictive models constitute significant research challenges. Most likely, there will be some synergy between language models used in speech and similar models that will be developed for low-level processing and correction of written ill-formed input.

Once the segmentation into basic units has been performed, the next step is to identify suitable lexical entries for each token and, in cases where more than one entry applies, to determine which one is most appropriate in the given context. This process is called part-of-speech disambiguation or POS tagging and is usually done with statistical models or machine-learning approaches trained on manually tagged data. Current technology achieves rather high accuracy on newspaper text, but again, performance suffers significantly when a model trained on a certain set of data is applied to text from a different domain. As the output of the POS tagger is typically used as input to subsequent modules, tagging errors may hamper the correct analysis of much more than the affected word. Research on high-quality POS tagging will face problems that are similar to those of language modelling: It requires detailed information about a large number of rare words that may be quite specific to the given domain and application, which is difficult to construct, no matter which road to lexical acquisition is taken. Any effort that will support the construction, distribution, sharing and re-use of large, domain-specific lexical resources will doubtlessly also help to improve the accuracy of POS tagging on text from these domains.

The next step in the analysis of text is to identify groups of words that belong together and refer to one semantic entity. Often, these phrases contain names, and for many practical applications, it is important to classify these expressions according to the type of entity they denote (Person, City, Company, etc.). Depending on the application, the classification may be more or less fine-grained. Again, it is obvious that improved lexical knowledge will help to improve the performance of named entity recognition. But we cannot in all cases rely on a lexical resource to cover the relevant entities. A text may discuss the opening of a new company, which will therefore not be contained in the lexicon. To handle such cases intelligently, we need mechanisms that can exploit contextual clues for the correct classification of unknown entities and we need effective mechanisms that propagate information about new entities into the lexical repositories, so that the system as a whole learns from the texts it sees, similar to the way a human reader would do.

Syntactic Analysis

The goal of syntactic analysis is to break down given textual units, typically sentences, into smaller constituents, to assign categorical labels to them, and to identify the grammatical relations that hold between the various parts.

In most applications of language technology the encoded linguistic knowledge, i.e. the grammar, is separated from the processing components. The grammar consists of a lexicon, and rules that syntactically and semantically combine words and phrases into larger phrases and sentences.

Several language technology products on the market today employ annotated phrase-structure grammars, grammars with several hundreds or thousands of rules describing different phrase types. Each of these rules is annotated by features and sometimes also by expressions in a programming language.

The resulting systems might be sufficiently efficient for some applications but they lack the speed of processing needed for interactive systems, such as applications involving spoken input, or systems that have to process large volumes of texts, as in machine translation.

In current research, a certain polarization has taken place. Very simple grammar models are employed, e.g. different kinds of finite-state grammars that support highly efficient processing. Some approaches do away with grammars altogether and use statistical methods to find basic linguistic patterns. Other than speed, these shallow and statistically trained approaches have advantages in terms of robustness, and they also implicitly perform disambiguation, i.e. when more than one analysis is possible, they make a decision for one reading (which of course may be the wrong one).

On the other end of the scale, we find a variety of powerful linguistically sophisticated representation formalisms that facilitate grammar engineering. These systems are typically set up in a way that all logically possible readings are computed, which increases the clarity (no magic heuristics hidden in procedures), but also slows down the processing. Despite their nice theoretical properties it has so far been difficult to adapt these systems to the needs of real-world applications, where speed, robustness, and partial correctness in typical cases are more urgent than theoretical faithfulness and depth of analysis.

How will this situation evolve? The two approaches will continue to compete for potential applications, and the current advantage for shallow approaches will diminish as more ambitious applications get within reach, and as languages are used that require richer analysis.

This will give incentives for shallow approaches to struggle for higher accuracy and more detailed analyses, whereas the deep processing will be forced to find workable solutions for the problems with speed and robustness. In the ideal case, more fine-grained forms of integration will be found, i.e. hybrid systems that will keep the advantages of both worlds as far as possible.

The simplest integration will just use shallow analysis as a fallback mechanism when deep analysis fails. In this case, results from both approaches need to be translated into one common representation, and the development of such a "common denominator" will be a significant challenge. To achieve an even more fine-grained cooperation between both approaches, deep analysis may be equipped with the ability to locally fall back to more superficial processing, driven by the need to deal with a specific problem in the input. Vice versa, the results of shallow analysis might be combined into a more detailed structure incrementally, based on rules from a deep grammar. Also analyses of corpus data obtained with shallow tools can be mined for linguistic knowledge that is then fed into resources used by a deep parser, and vice versa.

Research challenges will be how to find syntactic parsers that are at the same time fast, robust, deliver a detailed analysis that is correct with high probability and that are easily to adapt to special domains.

Semantic Analysis

The goal of semantic analysis is to assign meanings to utterances, which is an essential precondition for most applications of NLP. However, what level of abstraction is required in this phase depends on the difficulty of the task. Extraction of answers to simple factual questions from a given text will require less depth in analysis than the summarization of a lengthy treatise in few paragraphs.

We can dissect the task of semantic analysis into several subtasks, depending on the linguistic level where it takes place. Most important are the semantic tagging of ambiguous words and phrases, and the resolution of referring expressions.

The disambiguation of word senses needs to identify the meaning that should be assigned to a given word. The hardest part of this task is to define the set of meanings that should be considered in this task, i.e. to select the appropriate granularity for the conceptualization. The emergence of standardized, large-scale ontological resources will help to solve this part of the task, as the concepts that appear in such ontologies are a natural choice for the meanings of single words or simple phrases. Additionally, multilingual corpora that are aligned on the level of words and phrases can serve as an approximation to sense-tagged corpora, so draft ontologies and models for sense disambiguation can be extracted from these.

Considerable efforts in defining useful evaluation metrics for sense disambiguation are pursued in the ongoing SENSEVAL activities. So far, the methods used by the participants of SENSEVAL are mostly based on simple statistical classification using features extracted from the context of word occurrences. To the extent to which robust, high quality systems for syntactic analysis will appear, this will also help to obtain improved accuracy in the semantic disambiguation.

The resolution of referring expression such as pronouns or definite noun phrases is the ability to identify their target, which may be expressions that appear prior in the text, abstractions of material that appeared earlier, or entities that exist independently from the text in existing

background knowledge. Seen in a more general way, the task is to cull out objects and events from multimedia sources (text, audio, video). An example challenge includes extracting entities within media and correlating those across media. For example this might include extracting names or locations from written/spoken sources and correlating those with associated images. Whereas commercial products exist to extract named entities from text with precision and recall in the ninetieth percentile, domain independent event extractors work at best in the fiftieth percentile and performance degrades further with noisy, corrupted, or idiosyncratic data.

Therefore work on the resolution of referring expression and the identification of entities in text and multimedia documents remains important fields of activity for the future.

Discourse and Dialogue

Extracting the knowledge contained in documents and understanding and generating natural dialog behavior requires more than the resolution of local semantic ambiguities. Intelligent analysis needs to consider the global argumentative structure of documents and discourse, and dialogs need to be analyzed for pragmatic content.

Computational work in discourse has focused on two different types of discourse: extended texts and dialogues, both spoken and written, yet there is a clear overlap between these two: dialogues contain text-like sequences spoken by a single individual and texts may contain dialogues. But application opportunities and needs are different. Work on text is of direct relevance to document analysis and retrieval applications, whereas work on dialogue is of import for human-computer interfaces regardless of the modality of interaction. Both are divisible into segments (discourse segments and phrases) with the meaning of the segments being more than the meaning of the individual parts.

The main focus of the research is the interpretation beyond sentence boundaries, the intentional and informational approach.

According to the informational approaches, the coherence of discourse follows from semantic relationships between the information conveyed by successive utterances. As a result, the major computational tools used here are inference and abduction on representations of the propositional content of utterances.

According to the intentional approaches the coherence of discourse derives from the intentions of speakers and writers and understanding depends on recognition of those intentions.

One difficulty is to build models of human-machine-dialog when initially only examples of human-human interaction exist, which may not be relevant. Bootstrapping suitable models will therefore require Wizard-of-Oz studies with simulated systems.

Natural Language Generation

In many of the applications mentioned above, systems need to produce high-quality natural language text from computer-internal representations of information. Natural language generation can be decomposed into the tasks of text planning, sentence planning and surface realization. Text planners select from a knowledge pool which information to include in the output and out of this create a text structure to ensure coherence. On a more local scale, sentence planners organize the content of each sentence, massaging and ordering its parts.

Surface realizers convert sentence-sized chunks of representation into grammatically correct sentences.

Generator processes can be classified into points on a range of sophistication and expressive power, starting with inflexible canned methods and ending with maximally flexible feature combination methods. It is safe to say that at the present time one can fairly easily build a single-purpose generator for any specific application, or with some difficulty adapt an existing sentence generator to the application, with acceptable results. However, one cannot yet build a general-purpose sentence generator or a non-toy text planner. Several significant problems remain without sufficiently general solutions:

- Lexical selection is one of the most difficult problems in generation. At its simplest this question involves selecting the most appropriate single word for a given unit of input. However as soon as the semantic model approaches a realistic size and as soon as the lexicon is large enough to permit alternative locutions, the problem becomes very complex. The decision depends on what has already been said, what is referentially available from context, what is most salient, what stylistic effect the speaker wishes to produce and so on. What is required: development of theories about and implementations of lexical selection algorithms, for reference to objects, events states, etc., and tested with large lexical.

- Discourse structure (see also there) So far, no text planner exists that can reliably plan texts of several paragraphs in general. What is required: Theories of the structural nature of discourse, of the development of theme and focus in discourse, and of coherence and cohesion; libraries of discourse relations, communicative goals and text plans: implemented representational paradigms for characterizing stereotypical texts such as reports and business letters; implemented text planners that are tested in realistic non-toy domains.

- Sentence planning: Even assuming the text planning problem is solved, a number of tasks remain before well-structured multi-sentence text can be generated: These tasks, required for planning the structure and content of each sentence, include: pronoun specification, theme signaling, focus signaling, content aggregation to remove unnecessary redundancies, the ordering of prepositional phrases, adjectives, etc. What is required: Theories of pronoun use, theme and focus selection and signaling, and content aggregation; implemented sentence planners with rules that perform these operations; testing in realistic domains.

- Domain modeling: a significant shortcoming in generation research is the lack of large, well-motivated application domain models, or even the absence of clear principles by which to build such models. A traditional problem with generators is that the inputs are frequently hand-crafted, or are built by some other system that uses representation elements from a fairly small hand-crafted domain model, making the generator's inputs already highly oriented toward the final language desired….What is required: Implemented large-size (over 10.000 concepts) domain models that are useful both for some non-linguistic application and for generation; criteria for evaluating the internal consistency of such models; theories on and practical experience in the linking of generators to such models: lexicon of commensurate size.

Probably the problem least addressed in generator systems today is the one that will take the longest to solve. This is the problem of guiding the generation process through its choices when multiple options exist to handle any given input.

The generator user has to specify not only the semantic content of the desired text, but also its pragmatic – interpersonal and situational – effects. Very little research has been performed on this question beyond a handful of small-scale pilot studies. What is required: Classifications of the types of reader characteristics and goals, the types of author goals, and the interpersonal and situational aspects that affect the form and content of language; theories of how these aspects affect the generation process; implemented rules and/or planning systems that guide generator systems' choices; criteria for evaluating appropriateness of general text in specified communicative situations.

Effective presentations require the appropriate selection of content, allocation to media, and fine grained coordination and realization in time and space. Discovery and presentation of knowledge may require mixed media (e.g., text, graphics, video, speech and non-speech audio) and mixed mode (e.g., linguistic, visual, auditory) displays tailored to the user and context. This might include tailoring content and form to the specific physical, perceptual, or cognitive characteristics of the user. It might lead to new visualization and browsing paradigms for massive multimedia and multilingual repositories that reduce cognitive load or task time, increase analytic depth and breadth, or simply increase user satisfaction. A grand challenge is the automated generation of coordinated speech, natural language, gesture, animation, non-speech audio, generation, possibly delivered via interactive, animated lifelike agents. Preliminary experiments suggest that, independent of task performance, agents may simply be more engaging/motivating to younger and/or less experienced users.

Ontologies

Large-scale ontologies are becoming an essential component of many applications including standard search (such as Yahoo and Lycos), e-commerce (such as Amazon and eBay), configuration (such as Dell and PC-Order), and government intelligence (such as DARPA's High Performance Knowledge Base program). As discussed in the preceding paragraphs, ontologies will constitute a major source of knowledge needed for several levels of NLP.

Ontologies are increasingly seen as an important vehicle for describing the semantic content of web-based information sources and they are becoming so large that it is not uncommon for distributed teams of people to be in charge of the ontology development, design, population, and maintenance.

Ontologies define a vocabulary for researchers who need to share common understanding of the structure of information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them. The principal reasons to use an ontology in machine translation (MT) and other language technologies are to enable source language analyzers and target language generators to share knowledge, to store semantic constraints and to resolve semantic ambiguities by making inferences using the concept network of the ontology. An ontology contains only language independent information and many other semantic relations as well as taxonomic relations.

Though the utility of domain ontologies is now widely acknowledged in the IT (Information Technology) community, several barriers must be overcome before ontologies become practical and useful tools. One important achievement would be to reduce the time and cost of identifying and manually entering several thousand concept descriptions by developing

automatic ontology construction. Another important task is to find arrangements that make development and sharing of ontologies commercially attractive.

Some challenges for ontology research:

Work on ontologies needs to provide generally applicable top-ontologies that cover most important core concepts that will be needed for many domains. Extensions to new domains could then start by enriching these top-ontologies in a specific direction, reducing the initial effort for creating new ontologies, for merging independently developed extensions, and for rapid customisation of existing ontologies.

This requires that ontology-creators are willing to share parts of their work and find suitable processes to organize cooperation. It also requires the development of standards for the languages in which ontologies are specified and can be interchanged (e.g. along the lines of the OIL proposal). Here, the challenge is to find suitable compromises between expressive power and depth on one hand and ease of use on the other hand. Ideally, one specification language should be able to cover the whole spectrum up to advanced knowledge representation as used in the CYC project.

Incremental improvement of ontologies needs to be facilitated by specialized tools for easy visualization and modification. These tools (and the representations they work on) need to be domain-independent and suited even for casual users, and their design needs to be based on a user-centred process view.

It must be easy to plug in ontologies into various NLP-based tools such as tools for information extraction, organization and annotation of document collections (semantic Web), environments for terminology management and controlled language. This will permit to audit the contained knowledge in manifold ways, and will allow for rapid quality improvement.

What is required: tools that support broad ranges of users in (1) merging of ontological terms from varied sources, (2) diagnosis of coverage and correctness of ontologies, and (3) maintaining ontologies over time.

Lexicons

Lexical knowledge – knowledge about individual words in the language – is essential for all types of natural language processing. Developers of machine translation systems, which from the beginning have involved large vocabularies, have long recognized the lexicon as a critical (and perhaps the critical) system resource. As researchers and developers in other areas of natural language processing move from toy systems to systems which process real texts over broad subject domains, larger and richer lexicons will be needed and the task of lexicon design and development will become a more central aspect of any project.

A basic lexicon will typically include information about morphology and on the syntactic level, the complement structures of each word or word sense. A more complex lexicon may also include semantic information, such as a classification hierarchy and selectional patterns or case frames stated in terms of this hierarchy. For machine translation, the lexicon will also have to record correspondences between lexical items in the source and target language; for speech understanding and generation, it will have to include information about the pronunciation of individual words. For this purpose the overall lexicon architecture and the representation formalism used to encode the data are important issues.

No matter if we want to build an ontology or a lexicon, in general for this kind of high-quality semantic knowledge base, manual processing is indispensable. Traditionally computer lexicons have been built by hand specifically for the purpose of language analysis and generation. However, the needs for larger lexicons are now leading to efforts for the development of common lexical representations and co-operative lexicon development.

The area is ripe – at least for some levels of linguistic description – for reaching in the short term a consensus on common lexical specifications. We must expand the experiences with the sorts of semantic knowledge that could be effectively used by multiple systems. We must also recognize the importance of the rapidly growing stock of machine-readable text as a resource for lexical research. The major areas of potential results in the immediate future seem to lie in the combination of lexicon and corpus work. There's a growing interest from many groups in topics such as sense tagging or sense disambiguation on very large text corpora, where lexical tools and data provide a first input to the systems and are in turn enhanced with the information acquired and extracted from corpus analysis.

Machine Learning

As mentioned above, the acquisition of knowledge continues to impose on of the biggest difficulties to the application of NLP technologies. This holds both for linguistic knowledge (grammars lexicons) and for world knowledge (ontologies, facts). In order to make extensions of NLP to new domains possible, the acquisition process needs to be supported by algorithms that can exploit existing textual material and extract knowledge of various types from it.

Approaches to these methods can be found in various fields of research, such as statistical language models, bilingual alignment, grammar induction, statistical parsing, statistical classification technology, Bayesian networks and other ML methods used in artificial intelligence research, data mining techniques etc.

Due to the specific nature of lexical information, it is important to pick or develop methods that scale to large vocabularies and large sets of features and that can exploit multiple sources of evidence in a good way. Also, the methods need to be able to use a rich set of existing background knowledge, so that no effort is wasted in re-discovering what was already known.

It is important to have methods that can use richly annotated training data, but do not require that large datasets have to be annotated in this way. Instead, methods should be able to draw a maximum of advantage from raw data without annotation using unsupervised learning approaches. Also, it will be important to guide the effort of human annotation so that time is spent in the most efficient way, using active learning methods. Tools and processes for managing annotation projects (including assessment of quality levels) need to be developed and shared on a broad basis.

Whenever possible, one should try to use models that contain explicit linguistic representations (ideally organized along different strata) so that partial reuse of models and rapid adaptation to slightly different is facilitated.

4. Milestones

Some relevant items not included in Bernsen 2000.

Basic technologies

Short term

- accurate syntactic analysis for well-formed input from specific domains
- simple methods for minimizing annotation effort during domain adaptation
- ML algorithms that combine active and unsupervised learning for optimal exploitation of data
- generally applicable annotation schemes for semantic markup of text
- standards for encoding and exchange of ontological resources emerge
- top-level ontologies generally available
- tools for semi-automatic construction and population of ontologies from text
- tools for simple semantic enrichment of Web pages
- approaches to markup of discourse structure and pragmatics

Medium term

- improved methods for minimizing annotation effort during domain adaptation
- tools for adaptation of syntactic analysis to specific application with minimal human effort
- accurate syntactic analysis for slightly ill-formed input for restricted domains
- improved syntactic analysis of input with uncertainties (word lattices)
- machine learning methods that exploit and extend existing knowledge sources
- sufficiently accurate semantic analysis of free text from restricted domains
- generic schemes for the annotation of pragmatic content
- schemes for annotation of discourse and document structure
- generally usable ontologies exist for many domains
- NL generation verbalizes information extracted/deduced from multiple sources for QA
- Agent/user models for dialogs of moderate complexity

Long term

- accurate syntactic analysis for ill-formed input from multiple domains
- sufficiently accurate semantic analysis of free text from multiple domains
- recognition of pragmatic content in text and dialog
- NL generation produces stylistically adequate and well-structured text

Systems

Short term

- QA systems are able to answer simple factual questions
- Summarization system produce well-formed extracts from short documents
- automated e-mail response systems deliver high-quality replies in easy cases
- MT for information assimilation

Medium term

- QA systems that deduce answers from information in multiple sources
- Summarization systems are able to merge multiple documents
- Summarization systems are able to deliver different types of summaries
- Integration of translation memories with MT enables fast domain-adaptation
- Mixed-initiative dialogue systems for services and e-commerce

Long term

- Translator's workbenches based on TM, MT, and multi-modal input facilities
- QA systems that are able to explain their reasoning

5. Recommendations for NLP research in Europe

1. Build and make publicly available at low cost large-scale multilingual lexical resources, with broad coverage, generic enough to be reusable in different application frameworks
2. To turn special attention to the development of better ontologies which are reusable across domains in order to encode static world knowledge
3. Creation of large common accessible multilingual corpora of syntactical and semantically annotated data annotated also beyond sentence boundaries

4. Encourage development of statistical and machine-learning methods that facilitate bootstrapping of linguistic resources

5. Common standards will improve the effectiveness of people's cooperation, the identification of the requirements for the system specification, the inter-operability among systems and the possibility of re-using and sharing system components.

6. Integration of language processing into the rest of cognitive science, artificial intelligence and computer science e.g. some ambitious projects centered on NL but combining various techniques and different areas of AI. New type of projects: Very different for scale, ambition and timeframe

7. Establishment of centers of excellence as focus points for projects for a period of five to ten years.

8. Encourage systematic evaluations (but how ?)

6. References

- Berners-Lee, T. (2001) The Semantic Web, Scientific American (5/2001)

- Bernsen , N.O. (2000) Speech-Related Technologies. Where will the field go in 10 years? roadmap workshop, Katwijk

- Burger, J. e.a. (2000) Issues, Tasks and Program Structures to Roadmap Research in Question & Answering, Memo National Institute of Standards and Technology, Gaithersburg

- Carbonell, J. e.a. (2000) Vision Statement to Guide Research in Q&A and Text Summarization, Memo National Institute of Standards and Technology, Gaithersburg

- Cole,R.A. (Ed.). (1997) Survey of the State of the Art in Human Language Technology Cambridge University Press, Cambridge

- Declerck, Th., Wittenburg, P., Cunningham, H. (2001) The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment, ACL Workshop, Toulouse

- Delannoy, J.-F. (2001) What are the points? What are the stances? Decanting for question-driven retrieval and executive summarization, ACL Meeting, Toulouse

- Fensel, D. Hendler, J., Lieberman, H. ,Wahlster, W. (2000) Dagstuhl-Seminar: Semantics for the WWW, Dagstuhl, Germany

- Grishman, R. and Calzolari, N. Lexicons in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge

- Grosz, B. (1997) Discourse and Dialogue in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge

- Heisterkamp, P., (2000) Speech Technology in the year 2010, roadmap workshop, Katwijk
- Hirschman, L. and Thompson, H.S. (1997) Evaluation in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Hovy, E., (1997) Language Generation in Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Kang, S.-J. and Lee, J.-H. (2001) Semi-Automatic Practical Ontology Construction by using Thesaurus, Computational Dictionaries, and Large Corpora, ACL workshop Toulouse
- Kay, M. (1997) Machine Translation: The Disappointing Past and Present. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Kay, M. (1997) Multilinguality. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge
- Knight, K. (2001) Language Modeling for Good Generation, Workshop on Language Modeling and Information Retrieval, Pittsburgh
- Krauwer, St., (2000) Going from 'what' to 'why' across language barriers in the unified distributed information space. Roadmap workshop, Katwijk
- Maybury, M.T. and Mani, I., (2001) Automatic Summarization, ACL Meeting Toulouse
- Maybury, M.T., (2001) Human Language Technologies for Knowledge Management: Challenges and Opportunities, ACL Meeting, Toulouse
- Pardo, J.M., (2000) How will language and speech technology be used in the information world of 2010? Research challenges & Infrastructure needs for the next ten years. Report on the Roadmap Workshop, Katwijk aan Zee
- Staab, St., (2001) Knowledge Portals, ACL Meeting, Toulouse
- Stock, O. (2000) Processing Natural Language from 2000 to 2010, roadmap workshop, Katwijk
- Velardi, P. and Missikoff, M. and Basili, R. (2001) Identification of relevant terms to support the construction of Domain Ontologies, ACL workshop Toulouse
- Uszkoreit; H. (2001) Crosslingual Language Technologies for Knowledge Creation and Knowledge Sharing, Toulouse
- Zaenen, A. and Uszkoreit, H. (1997) Language Analysis and Understanding. In: Survey of the State of the Art in Human Language Technology, Cambridge University Press, Cambridge