Building the LR&E Roadmap

Joint COCOSDA and ICCWLRE Meeting

Preliminary Contributions for the Meeting

Working Material

Meeting Organisers

Stephan Busemann (DFKI, Saarbruecken)
Nicoletta Calzolari (ILC-CNR, Pisa)
Khalid Choukri (ELDA/ELRA, Paris)
Steven Krauwer (Utrecht University, Utrecht)

List of Contributors

- Sophia Ananiadou
- Steven Bird
- Sonja Bosch
- Paul Buitelaar
- Nicoletta Calzolari
- Jean-Pierre Chanod
- Fabio Ciravegna
- Kenneth Church
- Walter Daelemans
- Robert Dale
- Lars Degerstedt
- Christiane Fellbaum
- Dafydd Gibbon
- Eduard Hovy
- Shuichi Itahashi
- Arne Jönsson
- Gudrun Magnusdottir
- Wofgang Minker
- Carol Peters
- Roberto Pieraccini
- Stelios Piperidis
- Andrei Popescu-Belis
- Ehud Reiter
- Laurent Romary
- Justus Roux
- Florian Schiel
- Marc Schröder
- Gary Simons
- Kiril Simov
- Takenobu Tokunaga
- Inge Zwitserlood

Introduction

ELSNET is in the process of preparing a Technology Roadmap for Language and Speech Technology, to be combined with work done and being done by ELRA and ENABLER towards roadmaps for Language Resources and Evaluation. In our view a technology roadmap is a broadly shared vision of our future, which identifies the main technological challenges ahead of us, the prerequisites for addressing them and their potential impact in terms of applications or services they would enable.

Our approach is based on the definition of independent milestones (which could be resources, technologies or applications), and which are expected to be available in a specified year. Milestones are interlinked, as they can *require* or *enable* other milestones.

The ELSNET Roadmaps are available electronically and can thus easily be shared and updated. See http://elsnet.dfki.de, and click on "Resources", for our major example of a roadmap showing some advanced applications and the basic technologies and resources required for them.

On our website http://www.elsnet.org/roadmap.html you can find an account of our approach, an overview of the activities we have undertaken thus far (in particular with the support of the Enabler project, at a joint workshop in Paris on August 28-29 2003), and an initial overview of what we have, presented in a graphical format, to be found on http://elsnet.dfki.de.

The results we have collected during the various roadmap workshops and panel sessions we have organized are still being integrated, but at the same time we would like to start our process towards consensus building, as it is clear that a roadmap based on the opinions of a small group of people can not be seen as a broadly shared view of a whole community.

It is essential that more expert contributors get involved to better cover all areas, but also to feed us with more accurate facts and expectations (e.g. new approaches, new achievements and breakthroughs, etc.). The aim of the COCOSDA-ICCWLRE roadmap action is to arrive at a broadly supported update of the present roadmap, based on your contributions. We have asked many members from our community to provide us with their expert views with a view to integrating them into the roadmap in a harmonised way, consistent with other contributions. If there are any conflicts between the experts' views, we will use the meeting at LREC to discuss and resolve them.

Approach

We have asked a number of experts, preferably more than one for each subfield, whether they were willing to participate. If they agreed to participate in this task they were assigned a milestone, i.e. the very brief description of a specific technological goal that is indicated in the invitation letter we sent them. It was described in a very generic and global way, with a rather high level of ambition (e.g. "Machine Translation", or "Speech Understanding"). We see them as the major challenges for our field, and our experts were asked

(a) to tell us how (in their view) this goal could be made more specific and/or subdivided into subgoals that will bring us gradually closer to our ultimate goal, and that can be used to measure progress on our way forward, and
(b) corresponding to their decomposition of this goal, to complete a small template form (not more than one page) for each of the goals or subgoals they had identified.

The results will be incorporated in the present version of our roadmap, either by updating items that we already have, or by adding or deleting items. We have consulted a number of experts on each of the topics. As it is obvious that different experts may have different views, we will use the Roadmap Meeting we will be organizing at LREC 2004 in Lisbon (in the form of a Joint meeting of COCOSDA and ICCWLRE on May 30) to discuss emerging discrepancies with a view to create a consensus view.

In order to ensure the continuity of the roadmapping process in connection with language resources and evaluation we will propose that the responsibility for the process that should lead to continuous maintenance of the roadmap will be taken over jointly by COCOSDA and ICCWLRE. This could be implemented by means of organizing similar meetings at major conferences or by using other instruments.

For this action the following scenario was foreseen:

1. Experts propose sub-goals and corresponding descriptions for their area of expertise (due mid April)
2. We identify conflicts that require discussion and implement the undisputed parts in the roadmap (by early May)
3. We inform them about the result (agreements and disagreements) and ask them to prepare comments for the meeting at LREC on May 30) (by mid May); if they attend, they will be asked to present their comments, otherwise we will do it on their behalf.
4. Working meeting at LREC to build a consensus (May 30)

Task description, as given to the experts

For the field of expertise or milestone indicated in your invitation email you are invited
(a) (if necessary or desirable) to try to decompose the goal or challenge we gave you into further sub-goals that would eventually lead to the solution of the main problems; these sub-goals could be linearly ordered (e.g. "single speaker, isolated word speech recognition, small vocabulary", "speaker independent, isolated word recognition, small vocabulary", etc) or they could be part of a more structured hierarchy;
(b) to provide for each of the sub-goals some crucial data by means of the template below (a web version of the same form will be made available shortly at http://www.elsnet.org/roadmaptask.html); please note that we do not ask you to provide long prose documents, but rather just enough information to properly characterize the goal, its anticipated year of completion, the prerequisites and its potential impact.

The template form contains brief explanations of the type of information you are asked to provide. Please use one template form for each sub-goal and try to stick to the format.
The form is designed in such a way that your input should be easy to incorporate in our present roadmap (visit http://elsnet.dfki.de to see the graphical representation and click on a single item to see its description in tabular form). Please note that some of the questions in the table refer directly to other items already present in the web version of the roadmap, so we recommend that -in case you choose to use the MS Word template to complete the form- that you do it from a place where you have an internet connection.

An exercise of this type has necessarily uncertainties: technology evolves very fast, new technologies come in (from neighbouring or remote areas), markets can make 90 or even 180 degree turns, and completely new application areas can be opened up by new technological opportunities (e.g. the web, mobile communication). As a consequence it is very likely that in the end every prediction or expectation will turn out to be less accurate, than we would have hoped for, but at the same time it should be clear that (just like in traffic) we can only move forward if at any moment in time we have a concrete plan for our immediate and even longer term future, plus a willingness to reassess the situation as the world changes.

The Roadmap Task Force will collect the forms, compile the different contributions and integrate them in the roadmap, and put them on the agenda of the meeting at LREC, especially in cases where there is divergence between the experts' opinions.

At the meeting some of the milestones will be clustered to be presented together, while more debatable cases will be presented individually.

Template forms as given to the experts

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*** just your name*	*** your email*
Short name of the goal	*** just a short title to refer to it, e.g. "Multilingual Lexicon"*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*** one paragraph briefly describing the goal, in such a way that it is qualitatively and quantitatively verifiable whether it has been achieved, e.g. "A multilingual lexicon of ca 100000 entries for the 20 main languages, and good enough for machine translation with post-editing"*	*** same as above*
Expected year of completion	*** just a single year; if you would prefer a period, please reduce it to the middle year of the period; years as such are not the key issue here, but we need a simple instrument to put the challenges and milestones on a timeline*	*** same*
Justification	*** briefly indicate why you feel that this should be /would be achievable by the year you have given*	*** same*
Main obstacles for achieving the goal	*** main bottlenecks you see; this could include both technological and financial or organisational issues*	*** same*
Prerequisites	*** other technologies (tools, modules, systems) or language resources that do not yet exist and would enable or support this technology (please indicate which); please point to items already contained in our roadmap if applicable, but you can also add new ones if they are not present*	*** same*
Impact	*** other technologies or applications that would be enabled or supported (please indicate which) by this technology; please try to refer to items already included in the roadmap if possible, but feel free to add your own*	*** same*
Evaluation	*** one paragraph describing the approach to evaluation you think would be suited/needed for this*	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	
Milestone we asked you to describe	******** *as mentioned in the invitation email*
*** just a list of short names of sub-goals; for each of them we ask you to complete the sub-goal template form above*	
Comments	
*** whatever comments you have*	

CONTRIBUTIONS

Questionnaires received via the Web form

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 20 11:49:20 WET 2004
FORM ID: EhudReiter_goal_20040420114920
++
NAME: Ehud Reiter
EMAIL: ereiter@csd.abdn.ac.uk
++
YOUR GOAL:
Experimental evaluation methodology for NLG
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Devise an agreed methodology across the community for testing and evaluating NLG algorithms and choices with users. This is for user-based evaluation, not corpus-based evaluation. The methodology will include guidance on experimental design, appropriate controls, subject numbers, subject choice, statistical analysis, etc.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2006
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
We need robust and mutually agreed ways for evaluating different NLG algorithms and choices, and indeed for evaluating NLG as compared to other technologies for information presentation (such as visualisation). I think evaluation should be user-based, not corpus-based (see INLG-02 paper). Currently there is little discussion of experimental design or indeed "evaluation of evaluation", this needs to take place.
REFERENCES:
http://www.csd.abdn.ac.uk/~ereiter/papers/inlg02.pdf
++
YOUR OBSTACLES:
We need resources to run different types of experiments. We also need to get people to agree, which may not be easy. We may need to argue (discuss) with the rest of the NLP/speech community why we can't just use corpus-based evaluation like everyone else
REFERENCES:

++
YOUR PREREQUISITES:
none, this could start now
REFERENCES:

++

YOUR EXPECTED IMPACT:
Agreed evaluation techniques would be very helpful to NLG. Results would be easier to compare and understand, users of the technology could see how it was developing, etc.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Evaluation of evaluation techniques is hard, I'm open to suggestions!
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 20 12:00:17 WET 2004
FORM ID: EhudReiter_goal_20040420120017
++
NAME: Ehud Reiter
EMAIL: ereiter@csd.abdn.ac.uk
++
YOUR GOAL:
Empirical lexicons
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Create a lexicon which is empirically based on real language use,
including in particular its semantic component. This is a long-term
project, but an initial goal might be 1000 common words with relatively
simple meanings (such as "evening", "rising", or "above"). This could
be based on analysis of parallel text-data corpora (see reference),
although there are of course other techniques as well
REFERENCES:
http://www.csd.abdn.ac.uk/~ereiter/papers/lwm03.pdf
++
YOUR ESTIMATED YEAR OF COMPLETION: 2007
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
NLG systems need good lexicons, and systems which attempt to
communicate non-linguistic data in words need good models of how such
data maps to words.
REFERENCES:

++
YOUR OBSTACLES:
No real obstacles other than getting resources.
REFERENCES:

++
YOUR PREREQUISITES:
none, this could start now
REFERENCES:

++
YOUR EXPECTED IMPACT:
This is essential for data-to-text systems, which is my interest.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Use the methodology decided upon in my first subgoal
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 20 12:12:47 WET 2004
FORM ID: EhudReiter_goal_20040420121247
++
NAME: Ehud Reiter
EMAIL: ereiter@csd.abdn.ac.uk
++
YOUR GOAL:
Text Summaries of Complex Data
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Create a system which can generate textual summaries of complex numerical data. We have made a bit of progress towards this in our SumTime project (see reference), but this project has focused on relatively simple data. We need to look at more complex types of data, and also look at including qualitative inferences in generated summaries. This of course is vague, as a more concrete goal how about generating a summary of patient data from a hospital intensive care unit which is useful to a doctor, and another summary which is meaningful to the patient herself.
REFERENCES:
http://www.csd.abdn.ac.uk/research/sumtime/
++
YOUR ESTIMATED YEAR OF COMPLETION: 2009
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
I think this is a very promising application for NL Generation. The NLP community often talks about the world being flooded with text, but in fact the real flood is data, there is enormous quantity of sensor data being collected, most of which is currently ignored. If we can use language to communicate and summarise this data (and I suspect this will be especially useful to ordinary people who don't understand complex graphs), this will be very useful to society
REFERENCES:

++
YOUR OBSTACLES:
We need to talk seriously to people in data mining, HCI, etc, we cannot do this on our own. Unfortunately the NLP community currently seems fairly inward looking, and uninterested in what is happening elsewhere, which is a shame. In my experience data mining and HCI people are happy to talk as long as we're open-minded (and don't just try to "sell" NLP as the solution to all the world's problems). We'll also of course need some resources.
REFERENCES:

++
YOUR PREREQUISITES:
Evaluation methodology and empirical lexicon (my first two subgoals)
REFERENCES:

++

YOUR EXPECTED IMPACT:
As above, if we could do this well, this would be extremely useful technology for society/
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Use the methodology decided upon in my first subgoal
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 20 12:22:36 WET 2004
FORM ID: EhudReiter_goal_20040420122236
++
NAME: Ehud Reiter
EMAIL: ereiter@csd.abdn.ac.uk
++
YOUR GOAL:
Personal simplified web pages
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Create a web system which customised web pages according to the reading skill, background knowledge, etc of individual users. This is somewhat vague, a concrete initial application could be medical, such as a site which gave antenatal information to expectant mothers. This goal was also mentioned in the "Memories for Life" Grand Challenge developed by the UK Computing Research Committee.
REFERENCES:
http://www.csd.abdn.ac.uk/~ereiter/memories.html
++
YOUR ESTIMATED YEAR OF COMPLETION: 2014
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
This is a long-term project, which requires in particular detailed reader models, including knowledge of how to acquire such models and knowledge of how to use such models in NL Generation. I think 10 years is my best guesstimate, but this is really a guess
REFERENCES:

++
YOUR OBSTACLES:
Good understanding of reader models (as above), good understanding of how to map abstract information into words, etc. Lots of challenges, as this is a long-term goal!
REFERENCES:

++
YOUR PREREQUISITES:
lots, many of which are not on the current roadmap because they deal with user modelling and adaptation (which seems to be completely ignored in the current roadmaps)
REFERENCES:

++
YOUR EXPECTED IMPACT:
If we could do this well, it would be incredibly useful for society. In particular it might make a big difference to the lives of people from deprived backgrounds with limited skills and knowledge, who currently have a very hard time getting information,
REFERENCES:

++

YOUR EXPECTED EVALUATION NEEDS:
Use the methodology decided upon in my first subgoal
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Tue Apr 20 11:49:03 WET 2004
FORM ID: EhudReiter_list_20040420114903
++
NAME: Ehud Reiter
EMAIL: ereiter@csd.abdn.,ac.uk
++
YOUR TASK:
Generation
YOUR LIST OF SUBGOALS:
1) Experimental evaluation methodology for NLG
2) Empirical lexicons
3) Text summaries of complex data
4) Personal simplified web pages
++
YOUR COMMENTS:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 29 14:38:01 WET 2004
FORM ID: IngeZwitserlood_goal_20040429143800
++
NAME: Inge Zwitserlood
EMAIL: i.zwitserlood@viataal.nl
++
YOUR GOAL:
Extend grammatical descriptions of targeted sign languages
REFERENCES:
++
YOUR GOAL DESCRIPTION:
The knowledge of the grammatical structures of most sign languages is still (very) scant, usually focusing merely on particular aspects of the grammar, viz. those aspects that are different from (familiar) spoken languages.
REFERENCES:
++
YOUR ESTIMATED YEAR OF COMPLETION: 2015
REFERENCES:
++
YOUR JUSTIFICATION FOR THIS YEAR:
Linguistic sign language research in general focuses (deeply) on particular aspects of the sign languages in question; many aspects have hardly been topic of research yet

REFERENCES:
++
YOUR OBSTACLES:
A. There are only few resources for sign language research in many countries
B. Most sign language research focuses only on those grammatical aspects that are different from (familiar) spoken languages

REFERENCES:
++
YOUR PREREQUISITES:
A. Further development of good transcription methods
B. Training of sign language users as researchers; for most researchers the sign language they study is not their native language
REFERENCES:
A. http://www.mpi.nl/echo/tec-rep/wp2-tr14-2003.pdf
++
YOUR EXPECTED IMPACT:

REFERENCES:
++
YOUR EXPECTED EVALUATION NEEDS:

REFERENCES:
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 29 15:01:31 WET 2004
FORM ID: IngeZwitserlood_goal_20040429150131
++
NAME: Inge Zwitserlood
EMAIL: i.zwitserlood@viataal.nl
++
YOUR GOAL:
Recognition systems for sign languages
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Sign languages do not have a generally accepted writing system. Processing a sign language therefore needs recognition and interpretation of real-time signing. Systems that can capture signs have been developed but are still too crude for full recognition of the fine-tuned movements of the fingers. Furthermore, there are no systems yet that can interpret the captured signing.
REFERENCES:
Kennaway, R. (2002) Synthetic Animation of Deaf Signing Gestures. In: Lecture Notes in Computer Science Vol. 2298, pp. 146-157
++
YOUR ESTIMATED YEAR OF COMPLETION: 2020
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Especially for interpreting real-life signing extended linguistic knowledge of sign languages is necessary (but not present as yet).
REFERENCES:

++
YOUR OBSTACLES:
Grammatical description of target sign languages will take quite some time
REFERENCES:

++
YOUR PREREQUISITES:

REFERENCES:

++
YOUR EXPECTED IMPACT:

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:

REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 29 17:18:37 WET 2004
FORM ID: IngeZwitserlood_goal_20040429171837
++
NAME: Inge Zwitserlood
EMAIL: i.zwitserlood@viataal.nl
++
YOUR GOAL:
(Further) development of synthetic sign language rendering
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Currently there are two ways for rendering of synthetic signing by an avatar (computer animation): a. use of motion-captured signs and sign strings
b. use of genuine synthetic signs
Both need to be further developed.
REFERENCES:
Elliot, R. J.R.W. Glauert, J.R. Kennaway, I. Marshall (2000) The Development of Language Processing Support for the ViSiCAST Project. In: 4th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2000), pp. 101-108

Kennaway, R. (2002) Synthetic Animation of Deaf Signing Gestures. In: Lecture Notes in Computer Science Vol. 2298, pp. 146-157

Kennaway, R. (2004). Experience with and requirements for a gesture description language for synthetic animation. In: Lecture Notes in Computer Science Vol. 2915, pp. 300-311

++
YOUR ESTIMATED YEAR OF COMPLETION: 2008
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
There has been quite an amount of work in this field
REFERENCES:
Elliot, R. J.R.W. Glauert, J.R. Kennaway, I. Marshall (2000) The Development of Language Processing Support for the ViSiCAST Project. In: 4th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2000), pp. 101-108

Kennaway, R. (2002) Synthetic Animation of Deaf Signing Gestures. In: Lecture Notes in Computer Science Vol. 2298, pp. 146-157

Kennaway, R. (2004). Experience with and requirements for a gesture description language for synthetic animation. In: Lecture Notes in Computer Science Vol. 2915, pp. 300-311

++
YOUR OBSTACLES:
Not enough funding
REFERENCES:

++
YOUR PREREQUISITES:
Knowledge about the phonetics and phonology of sign languages
REFERENCES:

++
YOUR EXPECTED IMPACT:

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Native sign language users
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 29 17:34:03 WET 2004
FORM ID: IngeZwitserlood_goal_20040429173402
++
NAME: Inge Zwitserlood
EMAIL: i.zwitserlood@viataal.nl
++
YOUR GOAL:
(Perhaps) development and/or acceptance of a writing system for sign languages
REFERENCES:
++
YOUR GOAL DESCRIPTION:
Sometimes static representations of sign languages are needed. Existing notation systems are not accepted (yet) by the language users
REFERENCES:
Zwitserlood, I. & D. Hekstra (to appear) Sign Printing System - SignPS. In: Proceedings of LREC2004

http://www.signwriting.org/
++
YOUR ESTIMATED YEAR OF COMPLETION: 2015
REFERENCES:
++
YOUR JUSTIFICATION FOR THIS YEAR:
It will take some time before there will be enough knowledge about the phonology and phonetics of sign languages is needed.
Sign language users will have to accept a notation system developed for common language use. This will take some time.
REFERENCES:
++
YOUR OBSTACLES:
Insufficient knowledge about phonology and phonetics of sign languages.
Resistance of sign language users against the use of proposed notation systems.
REFERENCES:
++
YOUR PREREQUISITES:
Knowledge about the phonetics and phonology of sign languages

Acceptance by sign language user groups
REFERENCES:
++
YOUR EXPECTED IMPACT:

REFERENCES:
++
YOUR EXPECTED EVALUATION NEEDS:

REFERENCES:
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Thu Apr 29 17:26:46 WET 2004
FORM ID: IngeZwitserlood_list_20040429172646
++
NAME: Inge Zwitserlood
EMAIL: i.zwitserlood@viataal.nl
++
YOUR TASK:
Sign Language
YOUR LIST OF SUBGOALS:
Extended grammatical descriptions of targeted sign languages
Recognition systems for sign languages
(Further) development of synthetic sign language rendering
(Perhaps) development and/or acceptance of a writing system for sign languages

++
YOUR COMMENTS:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Mon Apr 26 16:31:25 WET 2004
FORM ID: JustusRoux_goal_20040426163124
++
NAME: Justus Roux
EMAIL: jcr@sun.ac.za
++
YOUR GOAL:
Telephone-based speech databases of official and/or widely used languages spoken in Africa
REFERENCES:
http://www.ast.sun.ac.za
++
YOUR GOAL DESCRIPTION:
Annotated limited domain telephone speech databases for official languages (of South Africa) and a widely used language in Africa (Swahili) of at least 15 to 20 hours of edited, usable speech in each language. Phonetic transcriptions of at least ten hours of speech is necessary to extract phonetic lexicons for each language.Limited domains to be determined by needs of communities,most probably related to health and social issues.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2007
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Given a lack of expertise and training of annotators (for specific languages), and given the experiences in the African Speech Technology (AST) project these are approximate (though realistic) dates.
REFERENCES:
Roux, JC, Louw, PH & Niesler, TR. (2004)The African Speech Technology project: An Assessment. ELREC 2004, Lisbon
++
YOUR OBSTACLES:
Financial support is probably the main obstacle. Suitably trained phoneticians will be required.
REFERENCES:
Roux, JC, Louw, PH & Niesler, TR. (2004)The African Speech technology project: An Assessment. ELREC 2004, Lisbon
++
YOUR PREREQUISITES:
Training of competent annotators/labelers for different languages is a prerequisite.Automated acoustic segmentation tools will enhance the work to be done.Add “Automatic segmentation tools (for speech)” in the “LanguageProcessing” category of the Roadmap.
REFERENCES:

++
YOUR EXPECTED IMPACT:
Speech based systems have the potential to bridge the digital divide in developing countries in an effective way, especially in illiterate communities. Although people may not be able to read or write, they

still have access to information through speech. This is supported by enormous growth in mobile telephony distribution in the African context.
Add “Speech-based information systems” in the “LangTech Applications” category of the Roadmap.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Either external validation of the speech databases (eg. by SPEX) or internally by means of a set of scripts developed for specific purposes.
REFERENCES:
Roux, JC, Louw, PH & Niesler, TR. (2004)The African Speech technology project: An Assessment. ELREC 2004, Lisbon
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Mon Apr 26 16:38:37 WET 2004
FORM ID: JustusRoux_goal_20040426163836
++
NAME: Justus Roux
EMAIL: jcr@sun.ac.za
++
YOUR GOAL:
Telephone-based speech databases of African varieties of English, French and Portuguese
REFERENCES:
http://www.ast.sun.ac.za
++
YOUR GOAL DESCRIPTION:
Annotated limited domain telephone speech databases for African varieties of English, French and Portuguese of at least 15 to 20 hours of edited, usable speech in each language. Phonetic transcriptions of at least ten hours of speech is necessary to extract phonetic lexicons for each language.Limited domains to be determined by needs of communities,most probably related to health and social issues.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2009
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Given a lack of expertise and training of annotators (for specific languages) - these are speculative dates:
2007 English as spoken in Central and East Africa;
2007 French as spoken in Central and West Africa;
2009 Portuguese as spoken in Angola and Mozambique

REFERENCES:

++
YOUR OBSTACLES:
Financial support and trained ’techno-linguists’ are probably the main obstacles.
REFERENCES:

++
YOUR PREREQUISITES:
Training of competent annotators/labelers for different languages is a prerequisite.
Automated acoustic segmentation tools will enhance the work to be done.
Add “Automatic segmentation tools (for speech)” in the “LanguageProcessing” category of the Roadmap.
REFERENCES:

++
YOUR EXPECTED IMPACT:
Speech based systems have the potential to bridge the digital divide in developing countries in an effective way, especially in illiterate

communities. Although people may not be able to read or write, they still have access to information through speech.
Add “Speech-based information systems” in the “LangTech Applications” category of the Roadmap.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Either external validation of the speech databases (eg. by SPEX) or internally by means of a set of scripts developed for specific purposes.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Mon Apr 26 16:18:24 WET 2004
FORM ID: JustusRoux_list_20040426161823
++
NAME: Justus Roux
EMAIL: jcr@sun.ac.za
++
YOUR TASK:
Speech databases of languages spoken in Africa
YOUR LIST OF SUBGOALS:
Limited domain annotated telephone speech databases of languages spoken in Africa. Two types resources are to be distinguished:(a) Official indigenous(African) languages and/or widely used languages in Africa (b) African varieties of, respectively, English, French and Portuguese
++
YOUR COMMENTS:
The development of language and speech technology applications in the African context is directly related to acceptance and uptake of technology in a specific country. This is largely determined by economic and/or socio-political factors.
Depending on technology uptake, speech databases for Swahili, as a major lingua franca in Africa should be developed.
Economic as well as socio-political development calls for the development of South-African English, Afrikaans and major official SA African languages (Xhosa, Zulu, Swati, Ndebele, Southern Sotho, Northen Sotho, Tswana, Venda and Tsonga).
Economic development (especially in the field of telecomunications) in Francophone countries calls for the development of African-French speech databases. (African-Portuguese may follow).
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Mon Apr 19 23:16:01 WET 2004
FORM ID: KennethChurch_list_20040419231600
++
NAME: Kenneth Church
EMAIL: church@microsoft.com
++
YOUR TASK:
a topic of your choice (to be agreed with us)
YOUR LIST OF SUBGOALS:

++
YOUR COMMENTS:
I would like to advocate the position in my Eurospeech-2003 keynote (see slides on http://research.microsoft.com/users/church/wwwfiles/publications.html conference paper 50). The question was: where have we been and where are we going. Some answers to this question are like Moore's Law where you can use historical progress to forecast future progress. Other answers are like the hockey stick business case where every year you promise to do great stuff next year. Slide 24 suggested that roadmap workshops have been exposed to the hockey stick criticism. I would like to push for a format where we would try to characterize progress in terms of a Moore's Law-like slope (e.g., error rates delining by an order of magnitude per decade).
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Sun Apr 11 16:28:01 WET 2004
FORM ID: LaurentRomary_list_20040411162800
++
NAME: Laurent Romary
EMAIL: Laurent.Romary@loria.fr
++
YOUR TASK:
Standards for Metadata
YOUR LIST OF SUBGOALS:
Subgoal 1: provide a stable infrastructure for the representation of metadata for language resources [emergency: 1-2 year]
Subgoal 2: get consensus on a core set of metadata descriptors for basic iidentification and management of language resources and tools (combining the experience gained from IMDI, OLAC and the TEI) [2 years]
Subgoal 3: implement a wider data category registry that integrates most descriptors used at present in language technology for a wide variety of languages [5 years]
Subgoal 4: expend the previous registry to become an archive of linguistic knowledge worldwide [10 years]
++
YOUR COMMENTS:
We should go towards a generalized notion of metadata for language resources that integrates both the classical view of those descriptors needed for the identification and basic documentation of language resources and tools, and the actual documentation of language resource content and structure (e.g. tagset associated to a POS annotation). This should allow our community to deploy an integrated semantic space of such descriptors (or data category, as defined in ISO committee TC 37).
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:49:55 WET 2004
FORM ID: RobertDale_goal_20040415034955
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
5. Shallow semantic summarisation
REFERENCES:

++
YOUR GOAL DESCRIPTION:
This follows on from subgoal #5; the aim here is to improve the quality of output that is possible by introducing a more sophisticated approach to the analysis of the source text, without yet pretending to achieve ‘full understanding’. The sense here is that the quality of summarisation will be improved if the text reconstruction mechanism has some idea of the meaning of the text, even if only at a superficial level. The major outcome here might be market leadership of a technology that improves upon the products deriving from subgoal #4, at least in some high-value domains.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2010
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Early results and prospects from subgoal #4 will provoke some teams to try to leapfrog the simpler technology.
REFERENCES:

++
YOUR OBSTACLES:
Any lack of perceived value from subgoal #4 will result in a shortfall in funding for targets such as this.
REFERENCES:

++
YOUR PREREQUISITES:
Better understanding of the shallow semantic requirements for generation.
REFERENCES:

++
YOUR EXPECTED IMPACT:
The likely development of a competing range of shallow semantic representations for text reconstruction.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
As for subgoal #4: For many text types, there are existing summaries that can serve as ‘gold standards’ for evaluation: for

example, we have abstracts in the case of academic papers and executive summaries in more business-oriented reports. A more general experimental framework can only be developed once there is a wider understanding of the needs of the consumer of the summary.

REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:51:12 WET 2004
FORM ID: RobertDale_goal_20040415035112
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
6. The development of a standardised architecture for adding natural language generation capabilities to relational databases
REFERENCES:

++
YOUR GOAL DESCRIPTION:
This follows on from subgoals #1 and #2: as we begin to see useful results in generating, for example, summaries of information in spreadsheets, more complex underlying datasets will begin to look worth attacking. We might expect the outcome here to be the provision of plug-ins by major database vendors such as Oracle that provide NLG reporting and summarisation functionalities for databases in a range of supported domains, probably based on the development of relevant XML-based standards.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2009
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Increasing provision of information by speech synthesis will drive this kind of technology forward.
REFERENCES:

++
YOUR OBSTACLES:
Difficulties in agreeing standard representation languages for use in databases.
REFERENCES:

++
YOUR PREREQUISITES:
Again here the major challenge is to identify a level of representation that is both transparent for database developers while providing the kind of information that makes it worthwhile using NLG techniques.
REFERENCES:

++
YOUR EXPECTED IMPACT:
Firm establishment of NLG as a component technology.

Likely development of a range of XML-based data description languages.

REFERENCES:

++

YOUR EXPECTED EVALUATION NEEDS:
Market evaluation.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:52:15 WET 2004
FORM ID: RobertDale_goal_20040415035215
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
7. Standardised mappings from widely used data formats to representations that can be used in NLG systems
REFERENCES:

++
YOUR GOAL DESCRIPTION:
This goal is in parallel to subgoal #6: while the vendors of databases will be interested in how they can make the contents of databases built on their platforms more accessible, the vendors of desktop office productivity applications will have a similar concern for their applications: imagine wanting to interrogate your Outlook schedule via the telephone in order to get a summary of what is happening in the week ahead. The major outcome here will be the development of a level of representation that can be used in conjunction with NLG technologies to provide such outputs; we might expect a vendor like Microsoft to settle on such a representation for its suite of desktop office applications.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2009
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
As for subgoal #6.
REFERENCES:

++
YOUR OBSTACLES:
The question of whether the major vendors of office applications will be willing to make the appropriate standards public.
REFERENCES:

++
YOUR PREREQUISITES:
A willingness for developers of NLG technologies to cater for input representations that are driven by application needs.

Involvement of NLG researchers in the appropriate standards developments.

REFERENCES:

++
YOUR EXPECTED IMPACT:
Increasing acknowledgement of the central role of NLG technologies in information applications.

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Market evaluation.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:53:10 WET 2004
FORM ID: RobertDale_goal_20040415035310
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
8. Multilingual generation services as part of the OS
REFERENCES:

++
YOUR GOAL DESCRIPTION:
This follows on from subgoal #7. As the benefit of NLG technologies here is appreciated and as the technology becomes better understood, we can expect to see the services required become part of the underlying operating system, whether this be on a phone, a PDA, a desktop computer, or some other as yet unseen platform. Outcome here is a widely understood NLG API that can be used by program developers to provide multilingual NLG reporting and output facilities in their applications.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2011
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
This development builds on a number of the other subgoals.
REFERENCES:

++
YOUR OBSTACLES:
It remains an open question as to whether a general purpose API that will work for a wide range of domains and applications will still be accessible to non-NLG experts.
REFERENCES:

++
YOUR PREREQUISITES:
The developments outlined in the previous subgoals.
REFERENCES:

++
YOUR EXPECTED IMPACT:
NLG firmly established as a component of information appliances.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Market evaluation.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:54:47 WET 2004
FORM ID: RobertDale_goal_20040415035447
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
1. The development of a standardised architecture for summarising tabular data structures in a specific domain
REFERENCES:

++
YOUR GOAL DESCRIPTION:
One of the most obvious areas where the linguistic sophistication of NLG techniques can be demonstrated is in the use of aggregation to provide concise descriptions of sets of similar or related facts. A common source of such facts is in tables; this goal is concerned with developing a standardised architecture and API that makes it possible to quickly and easily build components that can deliver natural language summaries of such data sources. The goal requires the development of an API that enables generation of texts from 80% of the simple tables that appear in a widely used domain, such as financial reporting. This would be likely to be available as a plug-in for a product such as Microsoft Excel.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2007
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
The basic NLG capabilities required here are already available; what is missing is the development of a standardised language for enabling their use. The ongoing development of standards such as XBRL provide a level of representation that should be able to support the generation task.
REFERENCES:

++
YOUR OBSTACLES:
The primary risk here is the possible lack of acceptance of the need to champion the task: if the NLG community does not see this as a relevant goal, then it will be taken up by others with different agendas, and the results may end up not taking account of valuable insights from the NLG community.
REFERENCES:

++
YOUR PREREQUISITES:
A better understanding of the nature and role of aggregation as an abstract process with respect to an arbitrary representation.

A simple ‘shallow realiser’ technology that makes it easy for non-experts to utilise NLG techniques: there have been some attempts at this but none that have yet been proven in real applications.

Active involvement of NLG proponents in the forums that define standards like XBRL.

REFERENCES:

++
YOUR EXPECTED IMPACT:
This technology would put NLG ‘on the desktop’, opening the door to other widespread uses of NLG technology.
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
As noted elsewhere, it is hard to see how the evaluation of NLG technology can be carried out meaningfully. The utility of a development such as that indicated in this subgoal would be judged by the market.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:55:43 WET 2004
FORM ID: RobertDale_goal_20040415035542
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
2. Extension of table summarisation to a wide range of domains and multiple languages
REFERENCES:
++
YOUR GOAL DESCRIPTION:
The success of the previous subgoal would provoke the development of similar technologies and techniques for other domains and languages, in each case occasioned by the availability of rich underlying resources such as we might hope to find on the semantic web. A general purpose solution here is unlikely, but an appropriate modularisation into domain-dependent and domain-independent components will arise through experimentation. This subgoal would likely result in tabular summarisation being available in five major European languages, plus Japanese and Mandarin, in three other high value domains.
REFERENCES:
++
YOUR ESTIMATED YEAR OF COMPLETION: 2008
REFERENCES:
++
YOUR JUSTIFICATION FOR THIS YEAR:
If people see the value in the results of the previous subgoal, we can expect many to jump on the bandwagon, with a consequent rapid development of the technology.
REFERENCES:
++
YOUR OBSTACLES:
Multiple efforts resulting in a plethora of interfaces.
REFERENCES:
++
YOUR PREREQUISITES:
Perceived success of subgoal #1.
REFERENCES:
++
YOUR EXPECTED IMPACT:
Further evidence that NLG has something to offer; acceptance of NLG as a component technology.
REFERENCES:
++
YOUR EXPECTED EVALUATION NEEDS:
As before: market evaluation.
REFERENCES:
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:56:48 WET 2004
FORM ID: RobertDale_goal_20040415035648
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
3. The development of a rich markup language that enables high level control of the prosody in text to speech
REFERENCES:

++
YOUR GOAL DESCRIPTION:
The goal here is something beyond standards like SSML, allowing both higher-level control of prosody that SSML provides, while also providing hooks that can be used appropriately by concept to speech systems by identifying the necessary and possible correlations between syntactic structure and prosody.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2007
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Much of the underlying theoretical work for this is probably already available.
REFERENCES:

++
YOUR OBSTACLES:
Lack of demonstration scenarios that convince both commercial and government sponsors to fund the work.

Difficulty in developing an agreed level of syntactic representation to act as a structure on which prosodic information can be overlaid.

REFERENCES:

++
YOUR PREREQUISITES:
Further improvements in TTS to demonstrate the utility of prosody.

Involvement of NLG researchers in the further development of standards such as SSML.

REFERENCES:

++
YOUR EXPECTED IMPACT:
Demonstration that NLG as a field has something to offer work in speech synthesis.

Improved multi-sentential TTS.

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
As before: market evaluation.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 03:57:59 WET 2004
FORM ID: RobertDale_goal_20040415035759
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR GOAL:
4. Syntactic smoothing of sentence-extraction based summarisation
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Existing commercially available summarisation techniques rely on simple sentence extraction. Coupled with some degree of broad coverage syntactic analysis, NLG makes it possible to produce smoother summaries by reconstructing sentences from parts of sentences. The major outcome here might be one or more products on the market that produce appreciably improved summaries of input documents.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2008
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Again, much of the preliminary research required to support this goal has been carried out; it is a question of putting together the pieces (and no doubt filling in some holes) in order to produce the required capabilities.
REFERENCES:

++
YOUR OBSTACLES:
Does anyone need or want text summarisation? Or if they do, do they want summarisation that is any better than that currently delivered by simple sentence extractors?

Lack of trust in automatically generated summaries.

REFERENCES:

++
YOUR PREREQUISITES:
Robust broad coverage parsing technologies that deliver structural analyses that can be used by a text reconstruction process.

Wider understanding of the nature and variety of summaries, particularly with respect to a user’s needs and how choices in this space can be communicated to the user.

REFERENCES:

++
YOUR EXPECTED IMPACT:

The text analyses required in order to support text reconstruction would be likely to be useful for other applications.

Development of a range of competing techniques for analysis-for-generation.

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
For many text types, there are existing summaries that can serve as ‘gold standards’ for evaluation: for example, we have abstracts in the case of academic papers and executive summaries in more business-oriented reports. A more general experimental framework can only be developed once there is a wider understanding of the needs of the consumer of the summary.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Thu Apr 15 03:43:56 WET 2004
FORM ID: RobertDale_list_20040415034356
++
NAME: Robert Dale
EMAIL: rdale@ics.mq.edu.au
++
YOUR TASK:
Generation
YOUR LIST OF SUBGOALS:
1. The development of a standardised architecture for summarising tabular data structures in a specific domain
2. Extension of table summarisation to a wide range of domains
3. The development of a rich markup language that enables high level control of the prosody in text to speech
4. Syntactic smoothing of sentence-extraction based summarisation
5. Shallow semantic summarisation.
6. The development of a standardised architecture for adding natural language generation capabilities to relational databases.
7. Standardised mappings from widely used data formats to representations that can be used in NLG systems.
8. Multilingual generation services as part of the OS.

++
YOUR COMMENTS:
It is tempting to specify subgoals in the area of generation with respect to the components of the now widely-accepted pipeline architecture for natural language generation (NLG) systems: discourse planning, sentence planning, and sentence realisation. One might target specific progress in each of these areas, perhaps in terms of ever-broader coverage. However, I believe this would be misleading as to the current state of the field. NLG is in the unfortunate position of still being a research area that delivers solutions that are looking for problems, and until we identify real problems where NLG can make a difference, it is very difficult to determine what a roadmap for the area might look like.

This may seem like a rather harsh position to take, and so I think it’s appropriate to offer here some argument in support of it.

Yorick Wilks is attributed with once noting that, if natural language understanding is like counting from 1 to infinity, then natural language generation is like counting from infinity to 1: a fundamental problem in natural language generation is thus the question of what you start from. Much work in generation proceeds in the following way: you identify some variation in surface form (it might be variations in syntactic forms that appear to convey the same underlying meaning; or variations in a text’s structure or content that appear to reflect the needs of different users while still being about the same topic); then, you hypothesise what underlying features might account for this variation (a notion of information structure in the first case above, or the parameters of a user model in the second); and then, you try to build a system that takes account of these features and their different combinations in order to build surface form variations of the type you were interested in. In so doing, you hope to explain the variations in terms of the underlying constructs. There is the separate tricky

question of how you evaluate such research (or NLG research in general), but I won’t try to address this question here.

This is fine in terms of a methodology for producing system fragments that can make ever finer linguistic distinctions, and it may indeed lead to the enrichment of linguistic theory. But this work is invariably carried out in a vacuum, devoid of a specific application that needs to make the distinctions explored.

This is not to say that there are no applications that might appear to require the generation of linguistic output. Superficially, at least, there are a number of such application areas we can point to:

• Spoken language dialog systems need to provide prompts and information to their users.
• Text summarisation systems need to produce summaries of input documents.
• Machine translation systems need to generate linguistic output in the target language.
• Grammar-checking systems need to produce corrected forms of sentences.

However, when we look at the current state-of-the-art in these areas, it becomes clear that NLG either does not have much to offer, or where it would appear to have something to offer, it is not being invited to offer it.

In the case of real spoken language dialog systems, current system output is invariably specified in the form of canned strings or simple templates. Of course, it is possible to argue that as the sophistication of dialog systems increases, richer generation capabilities will be required; but there is as yet no solid evidence that this is really the case.

Existing text summarisation technologies are still based on sentence extraction, with minimal reworking of the extracted material to provide fluency. Although there would appear to be scope for NLG in producing better summaries, the real bottleneck here is in the analysis of the original text to a level of sophistication that would enable such generation to take place.

Those working in machine translation are principally concerned with what corresponds to sentence realisation in the standard pipeline architecture: the mapping from some underlying representation of a sentence to its surface form. This is the one area where NLG research can most plausibly lay claim to having produced reusable resources (obvious examples are KPML and FUF/SURGE), but I am not aware of any MT systems that make use of those resources. Grammar checking systems are in the same position in this regard.

In essence, the generation community is at the stage where it can say to consumers, whether they be commercial or working in other areas of language technology research: We know how to generate such and such a range of phenomena automatically; all you have to do is provide input in the following form and we’ll do the rest. However, the consumers generally do not see the need for the range of output phenomena that can be delivered, and even if they do, the provision of

input in the required form is just too hard. If you want to build a multilingual system that summarises database content, the perception of the database developer is that it is easier to build simple templates in each of the target languages than it is to augment the database system with the required abstractions that would be needed to drive a generator. Until the NLG community can demonstrate real value to existing users and the applications they use, I think we are at an impasse with regard to the traditional research foci of NLG research. It is indeed possible that as underlying applications become more sophisticated, the need for NLG will become more apparent; but the jury is still out on that.

Of course, it’s not very helpful to suggest that we are stuck in the sand with no way forward, and I don’t in fact think that’s where we are. Rather, I think we have to see the future for the development of generation technologies mapped out in terms of incrementally adding capability to applications that exist today, or that can be expected in the medium term. The subgoals I have identified above, therefore, are derived from that perspective, rather than being derived from the smorgasbord of research topics that are investigated in the NLG community.

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 17:16:10 WET 2004
FORM ID: RobertoPieraccini_goal_20040415171610
++
NAME: Roberto Pieraccini
EMAIL: rpieracc@us.ibm.com
++
YOUR GOAL:
(Standards for metadata) requirement analysis
REFERENCES:

++
YOUR GOAL DESCRIPTION:
In this phase of the project we will collect the requirements for metadata annotation. For example
- Which type of data will the standard take into consideration (e.g. speech, text, ink, image, video, multimodal, etc.)
- What is the potential use of metadata in this context (e.g. automatic learning, documentation, inference, interactive systems, etc.)
- Types of metadata (e.g. reference, timing, alternative annotations, cross-document relationships, semantics, etc.)
- Relationships with existing recommendations (e.g. RDF, EMMA, etc.)
- Scope of the standard: at which level is the standard “normative”? (e.g. define an "ontology" or define a "format" for an ontology)

REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2005
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
This is a relatively easy task that can be based on other work as specified in the reference (e.g. EMMA). However we should not under evaluate the difficulties in predicting future needs .
REFERENCES:
http://www.w3.org/TR/EMMAreqs/

++
YOUR OBSTACLES:
Although there is some work we can rely on for speech and text, there is less on video and integrated multimodal data. Moreover, moving higher in the metadata abstractions (e.g. semantics) will certainly create a huge need to link with other resources (e.g. semantic web). However the main obstacle is accommodating and negotiating all the requirements needed by the community.
REFERENCES:

++
YOUR PREREQUISITES:

REFERENCES:

++

YOUR EXPECTED IMPACT:
The impact of this activity will reflect on the standard specification for metadata..
REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
In order to evaluate the requirements we should provide a limited list of case studies of system that would require metadata annotation. This can serve as a reference to see whether the requirement will satisfy the implementation of those systems..
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 17:52:42 WET 2004
FORM ID: RobertoPieraccini_goal_20040415175242
++
NAME: Roberto Pieraccini
EMAIL: rpieracc@us.ibm.com
++
YOUR GOAL:
(Standards for metadata) Specification document
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Given the set of requirements obtained with the previous sub-goal, we need now to come up with a description of the standard that will result in the specificatino of a markup language. The can be an extension of previous standard recommendations (such as EMMA) and based on existing paradigms (e.g. XML, RDF, ..)
REFERENCES:
http://www.w3.org/RDF/

http://www.w3.org/TR/emma/

++
YOUR ESTIMATED YEAR OF COMPLETION: 2006
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
The most reasonable way to proceed would be extend on existing specifications. If we do so, we can think of a first issue of a standard by 2006.
REFERENCES:

++
YOUR OBSTACLES:
We need to be able to cover the needs and evolution of the involved technologies. Again, the main obstacle in creating a standard specification is in the negotiation among different domains, sites, etc
REFERENCES:

++
YOUR PREREQUISITES:
Requirements
REFERENCES:

++
YOUR EXPECTED IMPACT:
The impact of a metadata standard is potentially on all the types of technology that require, produce or learn from annotations, either as their main objective or as an intermediate step. Examples of those are interactive multimodal systems with all various levels of multimodality (e.g. speech recognition, spoken language understanding, haptic interface, multimodal dialog, etc.) and information extraction and machine translation technologies in general.

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
The evaluation of a standard specification is done in several ways. In a first step we need to verify that the standard meets all the requirements. Then the standard is exposed to the public and feedback is collected. Finally, when reference implementations are available, we can proceed towards an analysis of the standard from a more practical point of view..
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Thu Apr 15 17:56:34 WET 2004
FORM ID: RobertoPieraccini_goal_20040415175634
++
NAME: Roberto Pieraccini
EMAIL: rpieracc@us.ibm.com
++
YOUR GOAL:
(Standards for metadata) Reference implementations
REFERENCES:

++
YOUR GOAL DESCRIPTION:
In order to fully validate a standard specification we need to have a few reference implementations that shows its functionality and effectiveness in meeting the initial requirements.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2006
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Implementation of simple proofs of concept can proceed along with the specification document, once it has reached a stable form.
REFERENCES:

++
YOUR OBSTACLES:

REFERENCES:

++
YOUR PREREQUISITES:
Specification document in a stable form.
REFERENCES:

++
YOUR EXPECTED IMPACT:

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
Reference implementations of the standard will be used to build prototypes that can be evaluated in a qualitative and quantitative way, depending on the applications.
REFERENCES:

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Thu Apr 15 17:07:23 WET 2004
FORM ID: RobertoPieraccini_list_20040415170722
++
NAME: Roberto Pieraccini
EMAIL: rpieracc@us.ibm.com
++
YOUR TASK:
Standards for metadata
YOUR LIST OF SUBGOALS:
1. Requirement analysis
2. Specification document
3. Reference implementations

++
YOUR COMMENTS:
Our requirements for HLT metadata annotation will continuously evolve through the course of the years; we cannot expect a corresponding standard to be a frozen specification, it needs to keep matching the evolving requirements in a few years from now. Thus standards, like technology, evolve continuously, and we have to take that into account in the compilation of the roadmap. However, we can consider here a first release of the standard, and keep in mind that further specifications will continue to evolve after that.
++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 27 12:14:57 WET 2004
FORM ID: SonjaBosch_goal_20040427121456
++
NAME: Sonja Bosch
EMAIL: boschse@unisa.ac.za
++
YOUR GOAL:
Written language corpora
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Multifunctional written language corpora of approximately 5 million words, which are shareable or available in the public domain, and which conform to international mark-up standards.
REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2006
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Written corpora of all the African languages spoken in South Africa (on average 5 million words per language), already exist. These corpora which are of a general nature, are mainly in plain text format with a minimal level of tagging.
An infrastructure for a national language and speech resource facility is in the process of being established by the Department of Arts and Culture, and should facilitate the re-usability and sharing of written corpora.

REFERENCES:
De Schryver Gilles-Maurice and DJ Prinsloo, 2000. The compilation of electronic corpora, with special reference to the African Languages, Southern African Linguistics and Applied Language Studies 18(1-4):89-106.

http://tshwanedje.com/tshwanelex/
http://www.dac.gov.za/about_us/cd_nat_language/language_planning/hlt_strategic_plan/hlt_strategic_plan2.htm
++
YOUR OBSTACLES:

 The high costs involved with creating linguistic resources.

 Willingness of research and academic institutions as well as companies to co-operate in efforts to centralise written language corpora in order to make them shareable or available in the public domain.

REFERENCES:

++
YOUR PREREQUISITES:

The adoption of common specifications and de facto international standards in creating written language corpora to ensure their compatibility at international and multilingual level

REFERENCES:

++
YOUR EXPECTED IMPACT:
Spelling checkers
Tokenisers
Morphological analysis
Disambiguation
Shallow parsing
Valuable data for statistical modelling and machine learning techniques
Evaluation of information retrieval systems

REFERENCES:

++
YOUR EXPECTED EVALUATION NEEDS:
The project LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards) aimed to pool together the European efforts of both academic and industrial actors towards the creation of de facto consensual standards for corpora, lexicons, speech data, and for assessing and evaluating resources.

The objective of the ISO/TC37/SC4 is to prepare international standards and guidelines for effective language resource management. This includes the development of principles and methods for creating, coding and processing of resources such as written corpora. Since the work also focuses on the evaluation of language resources, this would be an ideal approach to the evaluation of the written corpora in this subgoal.

REFERENCES:
http://www.hltcentral.org/cgi-bin/search-hlt.cgi?wm=wrd&m=all&q=EAGLES&submit=Search&np=0

Romary, Laurent & Nancy Ide. 2002. Standards for Language Resources. LREC 2002 Conference Proceedings, Vol 1. pp 59-65

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 27 12:19:18 WET 2004
FORM ID: SonjaBosch_goal_20040427121918
++
NAME: Sonja Bosch
EMAIL: boschse@unisa.ac.za
++
YOUR GOAL:
Machine-readable lexicons
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Machine-readable versions of published dictionaries in XML format containing approximately 30 000 entries, and which are shareable or available in the public domain, and which conform to international mark-up standards. Mono- and/or bilingual lexicons are included in the description.
Machine-readable specialist lexicons such as lexicons of proper names which include the most frequent surnames and first names.

REFERENCES:

++
YOUR ESTIMATED YEAR OF COMPLETION: 2007
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Published dictionaries of all the African languages spoken in South Africa already exist. National Lexicography Units for the official languages of South Africa are presently developing lexicons in electronic format.
An infrastructure for a national language and speech resource facility is in the process of being established by the Department of Arts and Culture, and should facilitate the development of machine-readable lexicons for shared use.

REFERENCES:
http://www.pansalb.org.za/index.php?nTab=7&lang_id=1

http://www.dac.gov.za/about_us/cd_nat_language/language_planning/hlt_strategic_plan/hlt_strategic_plan2.htm
++
YOUR OBSTACLES:

 Willingness of research and academic institutions as well as (publishing) companies to co-operate in efforts to make dictionary data available in machine-readable format for shared use.

 Although online dictionaries are reported on for some languages, they contain a maximum of 2000 to 3000 entries per language and do not include explicit linguistic information, which is a major disadvantage. In the case of Northern Sotho, however, a bilingual electronic dictionary SeDiPro 1.0 (de Schryver, 2003:10) containing over 20 000 entries with linguistic information, is available.

REFERENCES:
De Schryver, Gilles-Maurice. 2003. Online Dictionaries on the Internet: An Overview for the African languages. Lexikos: 13:1-20.

++
YOUR PREREQUISITES:
Consensus by developers of machine readable lexicons on common lexical specifications and de facto international standards to ensure their compatibility at international and multilingual level.
REFERENCES:

++
YOUR EXPECTED IMPACT:
The whole spectrum of language and speech technology, e.g.

Morphological analysis
Parsers and grammars
Shallow parsing
Semantic analysis
Machine translation etc.

REFERENCES:
Erjavec T, Evans R, Ide N and Kilgarriff A. 2003. From machine-readable dictionaries to lexical databases: the CONCEDE experience. Proceedings of COMPLEX 2003, 7th Conference on Computational Lexicography and Text Research, Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest. pp. 18-26.
++
YOUR EXPECTED EVALUATION NEEDS:
The project LRE-EAGLES (Expert Advisory Group on Linguistic Engineering Standards) aimed to pool together the European efforts of both academic and industrial actors towards the creation of de facto consensual standards for corpora, lexicons, speech data, and for assessing and evaluating resources.

The objective of the ISO/TC37/SC4 is to prepare international standards and guidelines for effective language resource management. This includes the development of principles and methods for creating, coding and processing of resources such as lexicons. Since the work also focuses on the evaluation of language resources, this would be an ideal approach to the evaluation of the XML lexicons in this subgoal.

REFERENCES:
http://www.hltcentral.org/cgi-bin/search-hlt.cgi?wm=wrd&m=all&q=EAGLES&submit=Search&np=0

Romary, Laurent & Nancy Ide. 2002. Standards for Language Resources. LREC 2002 Conference Proceedings, Vol 1. pp 59-65

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: SUBGOAL DESCRIPTION
SUBMITTED: Tue Apr 27 12:23:41 WET 2004
FORM ID: SonjaBosch_goal_20040427122341
++
NAME: Sonja Bosch
EMAIL: boschse@unisa.ac.za
++
YOUR GOAL:
Morphological analysers
REFERENCES:

++
YOUR GOAL DESCRIPTION:
Description:

Morphological analysers for computational analysis and synthesis of word forms.
The processing of the South African indigenous languages, which are characterised by complex morphological structures and are predominantly agglutinating in nature, particularly requires specialised tools for the automatic analysis of word-forms. Morphological analysis needs to be language-specific. The approach to developing morphological analysers can either be based on rules (finite-state grammars) and/or machine learning in order to partially automate the process.

REFERENCES:
Beesley KR and Karttunen L. 2003. Finite-state morphology. Stanford, CA: CSLI Publications.

http://www.fsmbook.com/
++
YOUR ESTIMATED YEAR OF COMPLETION: 2010
REFERENCES:

++
YOUR JUSTIFICATION FOR THIS YEAR:
Development of finite-state morphological analysers for four languages, namely Zulu, Xhosa, Ndebele and Northern Sotho, is already underway. The Zulu analyser prototype, which is closest to completion, will take approx. 4 years to complete. Therefore, given that machine-readable lexicons as basic resources become available in 2007, the development of analysers for the remaining languages could be fast-tracked in order to be completed in 2010.
Human capacity building, specifically in the field of computational morphological analysis is taking place by means of short, hands-on courses.

REFERENCES:
http://www.alasa.org.za/sig

http://www.conferences.hu/EACL03/Tut_WS.pdf
++
YOUR OBSTACLES:

 Human capacity - this is an interdisciplinary task involving linguists and computer programmers. There is no tradition of formal training of computational linguists in South Africa.

 Availability of machine-readable lexicons

REFERENCES:

++
YOUR PREREQUISITES:
Machine-readable lexicons as basic resources

Large corpora for automatic or semi-automatic discovery procedures that deduce rules and rule sets for morphological analysers

REFERENCES:
http://portal.acm.org/citation.cfm?id=637863&jmp=references&dl=GUIDE&dl=ACM

http://www.nisc.co.za/JournalHome/ling/abstracts/ling-v21-n4.htm#6
++
YOUR EXPECTED IMPACT:
Morphological analysis is the basic enabling application for further kinds of natural language processing, such as:

Lemmatising
Disambiguation
Shallow parsing
Semantic analysis
Machine translation
Document production
Information retrieval

REFERENCES:
Bosch SE and Pretorius L. 2002. The significance of computational morphological analysis for Zulu lexicography, in South African Journal of African Languages, 2002, 22.1:11-20.

++
YOUR EXPECTED EVALUATION NEEDS:
The finite-state calculus provides various powerful means of testing systems against large corpora, word lists, lexicons and lexical grammars.

REFERENCES:
Beesley KR and Karttunen L. 2003. Finite-state morphology. Stanford, CA: CSLI Publications.

http://www.fsmbook.com/

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Tue Apr 27 12:10:23 WET 2004
FORM ID: SonjaBosch_list_20040427121022
++
NAME: Sonja Bosch
EMAIL: boschse@unisa.ac.za
++
YOUR TASK:
Resources for minority languages
YOUR LIST OF SUBGOALS:

1. Written language corpora
2. Machine-readable lexicons
3. Morphological analysers

++
YOUR COMMENTS:
As agreed upon with Paola Baroni, I am focussing on written resources for minority African languages in the South African context. These languages are characterised by their highly agglutinating structures. Work on some of the subgoals identified above, has already begun in the case of certain languages, but has not been completed in most cases. Therefore for purposes of this submission, all the languages will be treated equally. In exceptional cases where work has been completed, it will be mentioned.

++
END OF THIS QUESTIONNAIRE
++

LREC ROADMAP QUESTIONNAIRE: LIST OF SUBGOALS
SUBMITTED: Tue Apr 13 10:12:24 WET 2004
FORM ID: StevenBirdandGarySimons_list_20040413101223
++
NAME: Steven Bird and Gary Simons
EMAIL: olac-admin@language-archives.org
++
YOUR TASK:
Standards for Metadata
YOUR LIST OF SUBGOALS:
1. Promote OLAC Metadata more widely within the Language Resources Community
2. Get community feedback on the OLAC controlled vocabularies
++
YOUR COMMENTS:
OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources. The OLAC Metadata Standard has recently been formally adopted, and is used by over two dozen language archives. OLAC search engines are hosted by LDC and LINGUIST.

++
END OF THIS QUESTIONNAIRE
++

Questionnaires received in MS Word format

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Paul Buitelaar*	*paulb@dfki.de*
Short name of the goal	*Ontologies*	
Description of the goal	*Ontologies are formal specifications of shared conceptualizations, representing concepts and their relations that are relevant for a given domain of discourse. Automation of ontology development (Ontology Learning) and use (Knowledge Markup; Ontology Population) can be implemented by a combination of linguistic analysis and machine learning approaches for text mining.*	*http://ontoweb-lt.dfki.de/*
Expected year of completion	*See below*	
Justification	*There will be many different levels in the application of this work, ranging from simple word/term frequency-based support in ontology engineering (already available), via linguistic/semantic analysis based support (some tools begin to emerge but some way to full integration into the ontology engineering process), up to nearly automatic ontology learning and population that will be fully integrated into semantic web applications ("distant" future: after 2010/2015 ?).*	
Main obstacles for achieving the goal	*Technological: accurate analysis of dependency structure (for many languages)* *Organizational: acceptance of text-based knowledge management tools and workflow*	
Prerequisites	*Fully integrated NLP grid/web*	

	services	
Impact	*Semantic Web; knowledge management*	
Evaluation	*Evaluation issues (quantitative and qualitative) will be discussed at the ECAI-04 workshop on Ontology Learning and Population (OLP). A guidelines report for the evaluation of these tasks will be compiled in the context of the workshop.* *In the context of the PASCAL NoE an evaluation task on OLP is expected to run over the next few years. A first sub-task (taxonomy extraction and population) is expected to run this year.*	*http://olp.dfki.de/ecai04/cfp.htm* *http://www.pascal-network.org/challenges/*

Template 1: description of sub-goals, 1 form for each sub-goal

Our question	***Your answer***	***References***
Your name	*Jean-Pierre CHANOD*	*chanod@xrce.xerox.com*
Short name of the goal 1	*"Expanding robust high-speed grammars with domain ontologies"*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*Recent progress in parsing and grammar writing led to high-speed broad-coverage parsers that mostly address the syntactic level, possibly enriched with shallow semantics (entity recognition, typing of selected relations). The goal is to develop language models and language resources able to bridge parsing and large-scale ontological representations, while preserving speed and robustness.*	*** same as above*
Expected year of completion	*** just a single year; if you would prefer a period, please reduce it to the middle year of the period; years as such are not the key issue here, but we need a simple instrument to put the challenges and milestones on a timeline*	*** same*
Justification	*** briefly indicate why you feel that this should be /would be achievable by the year you have given*	*** same*
Main obstacles for achieving the goal	*Language resource issues: developing large-scale reusable ontologies, challenge in mapping extra-linguistic ontological representations and language resource or models*	*** same*
Prerequisites		*** same*
Impact		*** same*
Evaluation		*** same*

Our question	*Your answer*	*References*
Your name	*Jean-Pierre CHANOD*	*chanod@xrce.xerox.com*
Short name of the goal 2	*"Expanding robust high-speed grammars at extra-sentential level: towards robust discourse analysis"*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*In recent years, robust parsers and associated grammatical descriptions mostly addressed the sentence level, while extra-sentential analysis focussed on specific sub-problems (e.g. pronominal coreference)* *The goal is to develop language models and language resources able to address extra-sentential relations with the same breadth and coverage as sentence parsing.*	*** same as above*
Expected year of completion	*** just a single year; if you would prefer a period, please reduce it to the middle year of the period; years as such are not the key issue here, but we need a simple instrument to put the challenges and milestones on a timeline*	*** same*
Justification	*Semantic interpretation based on currently available parsers focuses on local relations. Extracting global syntactic relations in conjunction with semantic interpreters will lead to more robust discourse analysis.*	*** same*
Main obstacles for achieving the goal	*** main bottlenecks you see; this could include both technological and financial or organisational issues*	*** same*
Prerequisites	*** other technologies (tools, modules, systems) or language resources that do not yet exist and would enable or support this technology (please indicate which); please point to items already contained in our roadmap if applicable, but you can also add new ones if they are not present*	*** same*
Impact	*** other technologies or applications that would be enabled or supported (please indicate which) by this technology; please try to refer to items already included in the roadmap if possible, but feel free to add your own*	*** same*
Evaluation	*** one paragraph describing the approach to evaluation you think would be suited/needed for this*	*** same*

Our question	*Your answer*	*References*
Your name	*Jean-Pierre CHANOD*	*chanod@xrce.xerox.com*
Short name of the goal 3	*"Multilingual Language resources"*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*As syntactic analysers will expand along the lines described above, the need to reach the same level of in-depth analysis across multiple languages will raise. This will require an extension and reinforcement around on-going activities in support of multilingual language resources, standards and evaluation.*	*** same as above*
Expected year of completion	*** just a single year; if you would prefer a period, please reduce it to the middle year of the period; years as such are not the key issue here, but we need a simple instrument to put the challenges and milestones on a timeline*	*** same*
Justification	*** briefly indicate why you feel that this should be /would be achievable by the year you have given*	*** same*
Main obstacles for achieving the goal	*The investment in developing and evaluating language resources will need to be supported by a clear view on the return of investment*	*** same*
Prerequisites	*** other technologies (tools, modules, systems) or language resources that do not yet exist and would enable or support this technology (please indicate which); please point to items already contained in our roadmap if applicable, but you can also add new ones if they are not present*	*** same*
Impact	*** other technologies or applications that would be enabled or supported (please indicate which) by this technology; please try to refer to items already included in the roadmap if possible, but feel free to add your own*	*** same*
Evaluation	*** one paragraph describing the approach to evaluation you think would be suited/needed for this*	*** same*

Our question	*Your answer*	*References*
Your name	*Jean-Pierre CHANOD*	*chanod@xrce.xerox.com*
Short name of the goal 4	*"Syntax in multimodal contexts "*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*Most effort in parsing, including robust parsing, focussed on somewhat normative texts (newspaper, technical documentation), while less normative language develops in every day's life (emails, phone, sms). The goal here is to develop specific parsing methodology to cope with robustness issues and support further accurate semantic interpretation.*	*** same as above*
Expected year of completion	*** just a single year; if you would prefer a period, please reduce it to the middle year of the period; years as such are not the key issue here, but we need a simple instrument to put the challenges and milestones on a timeline*	*** same*
Justification	*** briefly indicate why you feel that this should be /would be achievable by the year you have given*	*** same*
Main obstacles for achieving the goal	*** main bottlenecks you see; this could include both technological and financial or organisational issues*	*** same*
Prerequisites	*** other technologies (tools, modules, systems) or language resources that do not yet exist and would enable or support this technology (please indicate which); please point to items already contained in our roadmap if applicable, but you can also add new ones if they are not present*	*** same*
Impact	*** other technologies or applications that would be enabled or supported (please indicate which) by this technology; please try to refer to items already included in the roadmap if possible, but feel free to add your own*	*** same*
Evaluation	*** one paragraph describing the approach to evaluation you think would be suited/needed for this*	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	*Jean-Pierre CHANOD, chanod@xrce.xerox.com*
Milestone we asked you to describe	*Parsing or grammar*
Sub-goal 1 : *"Expanding robust grammars with domain ontologies"* Sub-goal 2: *"Expanding robust grammars at extra-sentential level: towards robust discourse analysis"* Sub-goal 3: *"Multilingual Language resources"* Sub-goal 4: *"Syntax in multimodal contexts"*	
Comments	
*** whatever comments you have*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Fabio Ciravegna*	*fabio@dcs.shef.ac.uk*
Short name of the goal	*Large scale Information extraction and integration*	*http://www.dcs.shef.ac.uk/~fabio/paperi/esws2004.pdf*
Description of the goal	*The Semantic Web needs annotated documents in order to make the semantics of documents available for automatic processing. Manual annotation is a bottleneck that is currently hindering the SemWeb realization. Information extraction and integration technologies should be provided in order to automatically produce large scale annotations for the Semantic Web. These annotation engines should work in a way similar to today's search engines constantly indexing documents with their semantics.*	*** same as above*
Expected year of completion	*2010*	
Justification	*Basic resources are currently under construction and they should be ready by 2006. At the same time the first prototypes of large scale annotation tools are emerging. I think by 2010 they should become common tools.*	*http://www.dcs.shef.ac.uk/~fabio/paperi/esws2004.pdf http://www2003.org/cdrom/papers/refereed/p831/p831-dill.html*
Main obstacles for achieving the goal	*Lack of a consistent community effort towards the goal so far (now it is changing with the UE 6th framework)*	
Prerequisites	*Technologies for:* 1. *automatic ontology learning* 2. *unsupervised learning for large scale information extraction* 3. *semi-supervised o unsuprevised technologies for large scale information integration* 4. *word sense disambiguation* 5. *human language technologies for the web (as opposed to for free texts)*	
Impact	*Creation of structured knowledge on a large scale*	
Evaluation	*There are formal ways of evaluating IE+II systems, but can be applied to limited scale*	

	evaluations (e.g. the MUC conference methodology). The dimension of the Web does not allow measuring easily a large scale IE+II task. Specific evaluation exercises are needed for this.	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Fabio Ciravegna fabio@dcs.shef.ac.uk***
Milestone we asked you to describe	***Information extraction***
Large scale information extraction and integration	
Comments	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Walter Daelemans*	*Walter.Daelemans@ua.ac.be*
Short name of the goal	*Acquisition and Learning (for NLP)*	*http://cnts.uia.ac.be/cnts/pdf/20040106.5156.dhdn03.pdf*
Description of the goal	*Automatic selection and optimization of Machine Learning (ML) methods for NLP tasks* ML will continue playing an important role in NLP for the development of robust and accurate NLP modules and applications. Several issues influence the success of a ML method applied to a task: the bias of the learning algorithm, the training sample and size, the feature selection and representation, the algorithm parameter settings, and **the interaction between all of these.** Powerful meta-learning methods will associate methods with the right bias for some task on the basis of the properties of the task (subgoal 1); powerful optimization techniques will provide models for tasks and applications with considerable higher accuracy and efficiency (subgoal 2).	
Expected year of completion	*2007*	
Justification	*Expected evolution of computing power*	
Main obstacles for achieving the goal	*Computing power*	
Prerequisites		
Impact	*All NLP modules and applications (accuracy and efficiency)*	
Evaluation	*Standard Machine Learning methodology for comparison, progress metrics etc.*	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Walter Daelemans***
Milestone we asked you to describe	***Acquisition and Learning***
1. Meta-learning methods helping in the selection and tuning of suitable ML methods (supervised, unsupervised, semi-supervised) on the basis of properties of the NLP task or application. *2. Optimization methods for sample selection, feature selection, algorithm parameter setting and interactions between all of these to build more optimal models for tasks.*	
Comments	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Sophia Ananiadou*	*S.Ananiadou@salford.ac.uk*
Short name of the goal	*Development of terminological resources*	
Description of the goal	*Monolingual terminological resources for a fast-growing discipline (e.g. genomics, molecular biology etc) based on automatic term extraction tools and existing controlled vocabularies. Application: aiding curation of scientific databases, semi-automatic ontology update, summarisation, IE, Q-A etc.*	*Ananiadou, S. Bodenreider, O. McGray, A Friedman, C. Cimino, JJ Zweigenbaum, P.*
Expected year of completion	*2005*	
Justification	*Terminologies backbone of data acquisition, knowledge management for specialised domains. Dynamic nature of biomedical areas demands systematic analysis of terminology. Consistent and up-to-date terminology required for many HLT applications, ie. IE, IR, QA. Sharing of resources for different types of applications requires agreement of standards.*	
Main obstacles for achieving the goal	*Term variation* *Domain dependence of term variation* *Scalability of a domain dependent technology to a wide variety of genres and text types* *Dealing with large data sets* *Acronym acquisition and generation* *Term ambiguity* *Term integration, linking terms from text to existing resources, data integration* *Tools for non experts*	*FASTR (Jacquemin, 2001)* *Hirschman (2002)* *Schwartz & Hearst (2003)* *Nenadic & Ananiadou (2003,2004)*
Prerequisites	*Term extraction tools* *Term management tools* *Named Entity Recognition tools* *Standards* *Term entry representation*	*C/NC (Frantzi, Ananiadou)* *LEXTER (Bourigault)* *TERMIGHT (Dagan & Church)*

		Standards *TBX* *OSCAR / LISA/ MARTIF, ISO TR 12618* *SALT (Budin, Melby)* *Galinski*
Impact	*For Human use, law, government, documentation, publishing, retrieval, eScience, Semantic Web* *Aiding to connect distributed information through common ontologies (based on terminology)* *For HLT applications IE systems, authoring, summarization, indexing, document characterisation, querying, question-answering.* *Annotation of texts* *Linking terminological information obtained from text to existing resources; automatic update of resources*	*Hahn, U. / Rector A. (medical terminology)* *Goble, C. (bio-ontologies)*
Evaluation	*Application dependent; lack of proper evaluation of term extraction tools*	*King, M.*

our question	*your answer*	*references*
Your name	*Sophia Ananiadou*	*S.Ananiadou@salford.ac.uk*
Short name of the goal	*Development of multilingual terminological resources*	
Description of the goal	*Multilingual terminologies for Cross Language Information Retrieval; support for minority languages*	*Hull, D.* *Grefenstette, G* *Ruch, P*
Expected year of completion	*2005*	
Justification	*Comprehensive terminologies for translation, knowledge transfer, ontologies, CLIR.* *Expansion to new EU languages.*	
Main obstacles for achieving the goal	*Term variation* *Scalability of a domain dependent technology to a wide variety of genres and text types* *Dealing with large data sets* *Term ambiguity* *New EU languages* *Tools for non experts*	*FASTR (Jacquemin, 2001)* *Ananiadou / Nenadic (2003)*
Prerequisites	*Term extraction tools* *Term management tools* *Standards* *Term entry representation*	*ACABIT (Daille)* *C/NC (Frantzi, Ananiadou)* *LEXTER (Bourigault)* *Standards* *TBX* *OSCAR / LISA/ MARTIF, ISO TR 12618* *SALT (Budin, Melby)* *Galinski*
Impact	*For Human use, law, government, documentation, (multilingual) publishing, retrieval, eScience, Semantic Web* *Access to knowledge across languages* *Aiding to connect distributed information through* *For HLT applications IE systems,(multilingual) authoring, summarization, indexing, machine translation, document characterisation, querying, question-answering, text generation*	*Buitelaar, P*
Evaluation	*Application dependent; lack of proper evaluation of term extraction tools*	*King, M.*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	Sophia Ananiadou / S.Ananiadou@salford.ac.uk
Milestone we asked you to describe	*Terminology*
Development of terminological resources Multilingual terminological resources Terminology management Standards for terminologies	
Comments	
Terminological resources are needed for tasks such as text mining, information extraction, information retrieval, machine translation, cross-language information retrieval. Tools for automatic term recognition (ATR) and term management (term variation, clustering and classification) are necessary for building resources. ATR is an enabler for knowledge acquisition and specification from concepts / terms. Discovery of relations and association between terms important for semi-automatic ontology building, update and maintenance.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Christiane Fellbaum*	*fellbaum@princeton.edu*
Short name of the goal	*"Multilingual Lexicon/Multilingual Lexicons that can intercommunicate"*	*http://www.globalwordnet.org*
Description of the goal	*"One multilingual lexicon, or many lexicons that are easily mappable of ca 30000 entries for ca. 20 main languages, and good enough for machine translation with post-editing"*	*same as above*
Expected year of completion	*Two years, depending of course on the financing*	
Justification	*Build on existing wordnets; some are more developed than others. Some are being created in critical languages but are not very large yet. Need time for the lexicography and esp. for setting the standards that everyone should follow to ensure compatibility. Consider some changes in the database design based on what we have learned in the past decade.*	*same as above*
Main obstacles for achieving the goal	*Most of the theoretical ideas are clear, but the building of the resources consumes much time and money. Organizational: put in place clear guidelines, with specific applications and goals in mind. Consider non-wordnet resources, too. Important: make resource freely available.*	
Prerequisites	*Synergy among many different groups. Sharing of experience. Medium-term stable financing.*	
Impact	*Enormous, esp. when including new languages wiht large user potential like Chinese, Hindi, etc.*	
Evaluation	*Intelligently designed lexical databases are needed for many applications; any or all of them can be used to evaluate the database.*	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Christiane Fellbaum/fellbaum@princeton.edu***
Milestone we asked you to describe	
Form a task group. *Agree on standards to follow for lexical database design.* *Implement standards rigorously; monitor development of databases.* *Periodic evaluation via applications.*	
Comments	
I've had time for a sketch only--let me know if we are thinking along the same lines.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Dafydd Gibbon*	*gibbon@spectrum.uni-bielefeld.de*
Short name of the goal	*Main goal:* *Resources for endangered languages*	*Ega model documentation <http://www.spectrum.uni-bielefeld.de/langdoc/>*
Description of the goal	*Provision of model resources for endangered languages with different typological characteristics (audio and video recordings, texts, transcriptions, annotations, sketch grammar, extended core lexicon) and appropriate acquisition tools.*	*OLAC <http://www.language-archives.org>*
Expected year of completion	*2010*	
Justification	*The date is optimistic. A number of model descriptive ventures are currently under way under the auspices of the EMELD, HRELP, DOBES and other projects, most of which, unfortunately, do not use state-of-the-art technologies.*	*EMELD <http://www.emeld.org>, HRELP <http://www.hrelp.org>, DOBES <http://www.mpi.nl/DOBES>*
Main obstacles for achieving the goal	*The main bottlenecks are connected with the "digital divide", i.e. the regrettably low priority of "commercially uninteresting" languages and societies with respect to infrastructural, educational and research funding. Specifically, the relatively tiny number of workers in this area compared with the large number of languages of the world (order of magnitude: 6000, most endangered) needs increasing by large-scale fundamental training schemes throughout the world.*	*International Clearing House on Endangered Languages <http://www.tooyoo.l.u-tokyo.ac.jp/Redbook/>, Endangered Langauge Fund <http://sapir.ling.yale.edu/~elf/>, Foundation for Endangered Languages <http://www.ogmios.org>*
Prerequisites	*A high priority should be the development of practical automated techniques for signal segmentation and annotation, and machine learning techniques for supporting lexicon acquisition and basic grammar induction. Likewise, open metadata portals are needed so that access to the*	*OLAC*

	data (subject to legal and ethical constraints) is maximally enabled.	
Impact	*Text-to-speech system development for use as information dissemination channels in pre-literate rural communities in minority and endangered language communities, as being developed by the Local Language Speech Technology Initiative..*	*LLSTI <http://www.llsti.org>*
Evaluation	*A complex of evaluation techniques is needed, both at the diagnostic level with regard to the resources themselves, and at the functionality level with regard to the utilization of resources for heritage preservation, language maintenance (for instance the development of language teaching materials) and scientific investigation.*	*Dafydd Gibbon, Roger Moore & Richard Winski, eds. (1997). Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter. Dafydd Gibbon, Inge Mertins, Roger Moore, eds. (2000). Handbook of Multimodal and Spoken Dialogue Systems. New York: Kluwer Academic Publishers.*

our question	*your answer*	*references*
Your name	*Dafydd Gibbon*	*gibbon@spectrum.uni-bielefeld.de*
Short name of the goal	*Audio and video recordings with transcriptions and annotations*	*Ega model documentation <http://www.spectrum.uni-bielefeld.de/langdoc/>*
Description of the goal	*Creation of new data in the field, or processing of legacy (analogue or digital) data.*	
Expected year of completion	*2006*	
Justification	*Model data are already available for some languages.*	*EMELD <http://www.emeld.org>*
Main obstacles for achieving the goal	*Not enough workers in the area to cope with the numbers of languages to cover, and with the expertise to produce transcriptions and annotations.*	
Prerequisites	*Provision of appropriate recording and computational equipment, and training in their use.*	
Impact	*Primary data for heritage preservation, language maintenance and scientific study.*	
Evaluation	*Evaluation according to accepted corpus design, production and processing techniques.*	*Dafydd Gibbon, Roger Moore & Richard Winski, eds. (1997). Handbook of Standards and Resources for Spoken Language Systems. Berlin: Mouton de Gruyter.* *Dafydd Gibbon, Inge Mertins, Roger Moore, eds. (2000). Handbook of Multimodal and Spoken Dialogue Systems. New York: Kluwer Academic Publishers.*

our question	*your answer*	*references*
Your name	*Dafydd Gibbon*	*gibbon@spectrum.uni-bielefeld.de*
Short name of the goal	*Audio and video recordings with transcriptions and annotations*	*Ega model documentation <http://www.spectrum.uni-bielefeld.de/langdoc/>*
Description of the goal	*Securing interpretability of legacy written text collections.*	
Expected year of completion	*2006*	
Justification	*Model text data are already available for some languages.*	*EMELD <http://www.emeld.org>*
Main obstacles for achieving the goal	*Not enough workers in the area to cope with the numbers of languages to cover.*	
Prerequisites	*Archiving of legacy text data.*	
Impact	*Primary data for heritage preservation, language maintenance and scientific study.*	
Evaluation	*Evaluation according to accepted corpus design, production and processing techniques.*	*EAGLES Written Corpus Working Group*

our question	*your answer*	*references*
Your name	*Dafydd Gibbon*	*gibbon@spectrum.uni-bielefeld.de*
Short name of the goal	*Construction of model sketch grammars for representative endangered languages.*	*Ega model documentation <http://www.spectrum.uni-bielefeld.de/langdoc/>*
Description of the goal	*Sketch grammars are generally constructed with traditional descriptive manual-intellectual techniques, using primitively formated word processor documents, whereas here comprehensive support in grammar structuring based on general questionnaires and on grammar induction is aimed at, with the aim of achieving greater efficiency in view of the large number of languages to be covered.*	
Expected year of completion	*2010*	
Justification	*Many traditional sketch grammars, and several questionnaires (effectively: ontologies) of grammatical categories are already available. A concerted effort would enable the creation of more systematic shared ontologies, such as the EMELD ontology GOLD (General Ontology for Linguistic Description).*	*EMELD <http://www.emeld.org>*
Main obstacles for achieving the goal	*Not enough workers in the area to cope with the numbers of languages to cover; basic and applied research needed to develop appropriate algorithms and data structures.*	
Prerequisites	*Corpora, grammar induction and "grammar workbench" tools.*	
Impact	*Basic components for TTS and other speech technology components to bridge the "digital divide" and the information technology gap.*	
Evaluation	*Evaluation according to accepted formalism design, production and processing techniques.*	*EAGLES Formalism Working Group*

our question	*your answer*	*references*
Your name	*Dafydd Gibbon*	*gibbon@spectrum.uni-bielefeld.de*
Short name of the goal	*Lexicon acquisition for representative endangered langauges.*	*Ega model documentation <http://www.spectrum.uni-bielefeld.de/langdoc/>*
Description of the goal	*Lexica are generally constructed with traditional descriptive manual-intellectual techniques, using primitively formated word processor documents, or with specialised tools such as Shoebox, or with spreadsheet software such as Excel, sometimes other database systems, whereas here comprehensive support in lexical class induction from corpora, and in the form of a sophisticated lexicographic workbench based on modern macrostructure, microstructure and mesostructure principles is needed, with the aim of achieving greater efficiency in view of the large number of languages to be covered.*	
Expected year of completion	*2008*	
Justification	*Semi-automatic lexicon development is relatively advanced, and sophisticated lexica could be created with proper training.*	*EMELD <http://www.emeld.org>*
Main obstacles for achieving the goal	*Not enough workers in the area to cope with the numbers of languages to cover.*	
Prerequisites	*Extensive text or transcribed corpora.*	
Impact	*Creation of dictionaries for heritage preservation, language maintenance and scientific study, and for language technology applications.*	
Evaluation	*Evaluation according to accepted lexicon design, production and processing techniques.*	*EAGLES Computational Lexicon Working Group*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Dafydd Gibbon <gibbon@spectrum.uni-bielefeld.de>***
Milestone we asked you to describe	***Resources for endangered languages***
Provision of model resources for endangered langauges with different typological characteristics: •audio and video recordings with transcriptions and annotations •texts, •sketch grammar, •extended core lexicon	
Comments	
See use of sub-goal template to describe main goal.	

Template 1: description of sub-goals, 1 form for each sub-goal

Our question	***Your answer***	***References***
Your name	*Gudrun Magnusdottir*	*esteam@otenet.gr*
Short name of the goal	*MachineTranslation – Text (Task)* Domain structure in language resources i.e. lexicons and texts	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	To promote research in Data Driven methods clear lines need to be made between areas in which they train for	*** same as above*
Expected year of completion	1 Year	*** same*
Justification	Data available and needs to be organised better. Some are already specified as such.	*** same*
Main obstacles for achieving the goal	No global domain structure theoretically viable thus the pragmatic approach of labelling the data with what comes to mind is the only choise. Keeping the resources clean is also very difficult.	*** same*
Prerequisites	prerequisites already exisiting but could need improvement	*** same*
Impact	statistical methods in general would be enhanced by being able to access structured data resources	*** same*
Evaluation	This can only be avaluated by practical use	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	
Milestone we asked you to describe	** *as mentioned in the invitation email*
*** just a list of short names of sub-goals; for each of them we ask you to complete the sub-goal template form above*	
Comments	
*** whatever comments you have*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Eduard Hovy*	*hovy@isi.edu*
Short name of the goal	*Cross Lingual Summarization: full summaries of mixed-language sources of different genres and domain/topics*	
Description of the goal	*Creation of the following collection:* • *a text, in various domains and genres (see details below)* • *at least two same-length summaries in each language, made by different humans (see language details below); if possible, also more, shorter or longer summaries,* • *for each summary group a score (or scores), produced by at least two different humans.* Each collection represents one combination as appropriate of (domain,genre), where • *Domain/topic = {news events, extended stories of events, travel/place descriptions, people/organization histories/bios},* • *Genre = {novels / films, email/bulletin board discussions, meeting transcripts, travelogues, biographies}.* *The more languages present, the better, but at least: English, one other European language, one Asian language, one more language (Arabic, Hindi, Chinese, etc. are of particular interest, given their sizes).* *Ideal amounts: at least 250 texts in each domain/genre combination.*	
Expected year of completion	*2006*	
Justification	*This collection should be built in stages. Parts of it (news) can simply be assembled from existing DUC and other resources, and can be ready in a few months, after translation. Other parts (novels and bboard discussions) can be bought and/or downloaded, with summaries, and also need translation, but since they are more complex summarization tasks, their scoring still has to be performed. For these I expect 12 to 18*	

	months after collection initiation. For yet others, such as meeting notes, the task is quite unknown and just producing summaries, and then scoring and translating them, will take perhaps 2 years.	
Main obstacles for achieving the goal	*Financial: to pay people to produce and translate the summaries.* *Methodological: for some of the summary types, some early investigation is required to determine scoring methods.*	
Prerequisites	*No significant ones.*	
Impact	*Information Extraction, question answering (with complex answers), and possibly in a small way MT*	
Evaluation	*Intrinsic evaluation: automatically with ROUGE, and manually the normal DUC way using the SEE interface.* *Extrinsic evaluation: The task of multilingual report writing. Given a summary (vs. the full text, or vs. a summary in another language), create a report as specified. The report is manually scored for content, coherence, etc.*	*ROUGE papers by Lin and Hovy in recent conferences and DUC workshops*

our question	*your answer*	*references*
Your name	*Eduard Hovy*	*hovy@isi.edu*
Short name of the goal	*Cross Lingual Summarization: headline summaries of mixed-language sources*	
Description of the goal	*Creation of the following resource:* *A collection of texts in various source languages, each with at least two (and hopefully four) headline-length summaries in (at least) the following languages, made by different humans: English, one other European language, one Asian language, one more language (Arabic, Hindi, Chinese, etc. are of particular interest, given their sizes). Together with each such headline, scores (at least 2, hopefully four) made by independent multilingual humans.* *Ideal amounts: at least 1000 texts in four languages.*	
Expected year of completion	*2006*	
Justification	*It's just a matter of doing it. I estimate 10 to 15 per person per hour, that's about 1 month for 1000 texts by one half-time person. Hire 4 summarizers for a year and 4 scorers for 3 months and in 15 months there is a corpus of 12000 texts, each with four (plus original) headlines, scores 4 times, for each language. Half this amount for two languages, one third for three, etc.*	
Main obstacles for achieving the goal	*Financial: to pay people to produce and translate the summaries.*	
Prerequisites	*No significant ones.*	
Impact	*Information Extraction*	
Evaluation	*Intrinsic evaluation: automatically with ROUGE, and manually the normal DUC way using the SEE interface.* *Extrinsic evaluation: IR relevance judgments*	*For intrinsic tests: ROUGE papers by Lin and Hovy in recent conferences and DUC workshops. For extrinsic test: forthcoming paper by Zajic, Schwartz, and Dorr (Maryland)*

our question	*your answer*	*references*
Your name	*Eduard Hovy*	*hovy@isi.edu*
Short name of the goal	*Cross Lingual Summarization: summaries of multi-document mixed-language sources*	
Description of the goal	*Creation of the following resource:* *A collection of sets of texts, each set devoted to a single topic, in a single genre. But each set contains at least to (and up to four) different languages (including English, one European, and one Asian language). With each set, at least two (and hopefully four) paragraph-length summaries, made by different humans, in at least English, but possibly also in one other European language. With each summary, scores (at least 2, hopefully four) made by independent multilingual humans.* *Ideal amounts: at least 500 topic collections.*	
Expected year of completion	*2007*	
Justification	*This is more work tan headline creation, but a similar time/effort computation holds.*	
Main obstacles for achieving the goal	*Financial: to pay people to produce and score the summaries.*	
Prerequisites	*No significant ones.*	
Impact	*Information Extraction, machine translation, IR*	
Evaluation	*Intrinsic evaluation: automatically with ROUGE, and manually the normal DUC way using the SEE interface.* *Extrinsic evaluation: IR relevance judgments*	*For intrinsic tests: ROUGE papers by Lin and Hovy in recent conferences and DUC workshops. For extrinsic test: forthcoming paper by Zajic, Schwartz, and Dorr (Maryland)*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	*Eduard Hovy, hovy@isi.edu*
Milestone we asked you to describe	*Achievement milestone:* • *Creation of each individual resource* *Dependency milestones:* • *Identification of suitable source text collection in each domain/genre* • *Definition and testing of suitable evaluation metric for each domain/genre*
Resource 1: full summaries of mixed-language sources of different genres and domain/topics *Resource 2: headline summaries of mixed-language sources* *Resource 3: summaries of multi-document mixed-language sources*	
Comments	
Since each resource should contain human evaluation scores, the collection process should be carefully coordinated with evaluation specialists.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*** Shuichi ITAHASHI*	*** itahashi@is.tsukuba.ac.jp*
Short name of the goal	*** Multilingual parallel speech corpus*	***S. Itahashi et al, "Design and Creation of Multilingual Speech Corpus," Proc. SNLP-Oriental COCOSDA 2002, Hua Hin, Thailand, pp. 49-53 (May, 2002)*
Description of the goal	*** Multilingual parallel speech corpus of 100 or 200 basic words and 500 phonetically-rich sentences for 30 main languages to be used for phonetic/phonological analysis of language similarity*	*** same as above*
Expected year of completion	***2007*	*** same*
Justification	*** It will take about a few years to collect the speech material and a few more years for investigating the similarity of languages.*	*** same*
Main obstacles for achieving the goal	*** 1) Automatic method of segmenting speech of various languages into phonemic units.* *2) Organization of collecting multilingual parallel speech corpus.*	*** same*
Prerequisites	*** Language identification methods, distance measures between two spoken languages.*	*** same*
Impact	*** It will become possible to make clear the similarity among various languages based on speech data including those languages which do not have letters or transcription systems.*	*** same*
Evaluation	*** comparison with the dendrogram or tree structure of language families already known*	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Shuichi ITAHASHI:itahashi@is.tsukuba.ac.jp***
Milestone we asked you to describe	******** *Speech Resources*
**	
Comments	
**	

Template 1: description of sub-goals, 1 form for each sub-goal

Our question	*Your answer*	*References*
Your name	*Arne Jönsson & Lars Degerstedt*	*arnjo@ida.liu.se*
Short name of the goal	*Evolutionary development of dialogue systems*	*http://www.ida.liu.se/~arnjo/papers/johansson-d-j.pdf*
Description of the goal	*A language engineering framework for evolutionary development of dialogue systems. To identify, conceptualise, design and implement domain-independent facility software and domain-dependent sample applications that incorporates dialogue capacities. The choosen strategy should support ease-of-use and ease-of-development for both concepts and software.*	*** same as above*
Expected year of completion	*The evolutionary approach means that we see no final year. A conceptual foundation for an evolutionary framework. can be ready by 2005 handling basic information-providing dialogue systems that allows for continous development*	*** same*
Justification	*Implementation of component-based, reusable and effectively engineered mixed-initiave dialogue systems is to be done in an evolutionary fashion.. From the experience gained developing various dialogue systems new knowledge arises.*	*** same*
Main obstacles for achieving the goal	*The role of software engineering for natural language processing is unclear and not recognised enough, within the research communities. By software engineering, we here understand such activities and results as software design and construction, methodology, and learning from experiences of finished software projects.*	*** same*
Prerequisites	*Generic facility software for various dialogue tasks such as dialogue history management and dialogue control, suitable as a starting point for evolutionary refinement.*	*** same*
Impact	*Filling the gap between approaches and agendas to development of dialogue systems in the indusrty and the research communities,.*	*** same*
Evaluation	*The ease and effectiveness of using the framework for development of robust dialogue systems, from an engineering point of view.*	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	*Arne Jönsson & Lars Degerstedt / arnjo@ida.liu.se*
Milestone we asked you to describe	******** *as mentioned in the invitation email*
*** just a list of short names of sub-goals; for each of them we ask you to complete the sub-goal template form above*	
Comments	
*** whatever comments you have*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	Wolfgang Minker	wolfgang.minker@e-technik.uni-ulm.de
Short name of the goal	Creation and Availability of Behavioral Data Resources	*Knudsen et al. (2002): Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. ISLE Deliverable D8.1*
Description of the goal	*Create and study re-usable facial, gesture or bodily posture data resources with or without accompanying speech.*	
Expected year of completion	*2009*	
Justification	*Gesture as well as facial data resources are already available to some extent. Substantial data collection effort is required for bodily posture data.*	
Main obstacles for achieving the goal	*Resources are usually created for specific application purposes and may not easily be re-used for other domains and modality combinations.*	
Prerequisites	*Availability of data annotation tools and schemes.*	
Impact	▪ *Facilitates multimodal spoken language dialogue systems specification and development.* ▪ *Enables evaluation of multimodal spoken language dialogue systems.* ▪ *Data studies enhance systems usability.* ▪ *Availability of re-usable data reduces system development costs.*	
Evaluation		

our question	*your answer*	*references*
Your name	Wolfgang Minker	wolfgang.minker@e-technik.uni-ulm.de
Short name of the goal	Uniform Data Annotation Tools and Schemes	
Description of the goal	*Create standardized tools supporting the annotation of spoken dialogue, facial expression, gesture or bodily posture data. Perform this annotation according to specific coding schemes to be specified for all relevant classes of behavioral phenomena involved in multimodal interaction.*	*Bernsen et al. (2003): Best practice in natural and multimodal interactivity engineering. CLASS Deliverable D1.5+6*
Expected year of completion	*2008*	
Justification	*Several projects dealing with the creation of annotation tools mention standardization as a goal.*	
Main obstacles for achieving the goal	• *Robustness, stability and real-time performance problems of the tools.* • *Variety of possible semantic and dialogic representations on the higher language levels and for non-speech data.*	
Prerequisites	• *Availability of a sufficient amount of expressive multimodal data resources.* • *Involvement of industry to generate stable and product-like annotation software tools.*	
Impact	• *Make transcription, annotation and data analysis considerably more efficient compared to a completely manual process.* • *Facilitate and reduce the cost of production and exploitation of data resources.*	
Evaluation		

our question	*your answer*	*references*
Your name	Wolfgang Minker	wolfgang.minker@e-technik.uni-ulm.de
Short name of the goal	Common Multimodal Spoken Language Dialogue Systems Development and Evaluation Platforms	*http://fofoca.mitre.org/*, *http://www.corba.org/*, *http://www.ai.sri.com/~oaa/*, *http://www.w3.org/*
Description of the goal	*Create re-usable platforms, components and system architectures, development toolkits, interface languages, data formats and standards.*	
Expected year of completion	*2009*	
Justification	*Transatlantic and national European efforts to coordinated projects already exist.*	*Pallett et al. (1994): 1994 Benchmark tests for the ARPA spoken language program, ARPA SLT Workshop.* *Mariani et al.(1999): Human language technologies evaluation in the European framework, DARPA Broadcast News Workshop.* *http://communicator.sourceforge.net/*
Main obstacles for achieving the goal	▪ *Interdisciplinary character of the different technologies involved makes this task considerably complex.* ▪ *Unlike in the US, working on a common task using common data and development platforms has not been a clear focus of European programs yet supporting diversity of research.*	*Mariani (1998): Evaluating Evaluation: US vs EU, ELSNews 7.8*
Prerequisites	▪ *Standardization of data annotation schemes.* ▪ *Availability of a sufficient amount of expressive multimodal data resources.* ▪ *Substantial funding and co-ordination of competitive international evaluation projects.*	
Impact	*Enable developers an easy access to highly performant system components which are not in the development focus.*	
Evaluation		

our question	*your answer*	*references*
Your name	Wolfgang Minker	wolfgang.minker@e-technik.uni-ulm.de
Short name of the goal	Usability Evaluation Standards for Multimodal Spoken Language Dialogue Systems	*Dybkjær et al. (2004): Usability Evaluation of Multimodal and Domain-Oriented Spoken Language Dialogue Systems, LREC.*
Description of the goal	*Evaluate the appropriateness of the proposed interaction modalities in relation to the application and the targeted user group.*	
Expected year of completion	*2010 or later*	
Justification	*Usability evaluation standards for unimodal spoken language dialogue systems have not yet been established.*	
Main obstacles for achieving the goal	▪ *Definition of criteria for evaluating the combinatorial contribution to usability and user satisfaction of the non-speech input and/or output modalities.* ▪ *Usability evaluation of unimodal spoken language dialogue systems is still only baseline.*	
Prerequisites	▪ *Existing usability evaluation baseline of unimodal spoken language dialogue systems may in part be re-used.* ▪ *Decision, what to transfer from this baseline and which new criteria and metrics are required.* ▪ *Additional user needs analyses need to be carried out.*	*Gibbon et al. (1997): Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, New York.* *Walker et al. (1997): PARADISE: A Framework for Evaluating Spoken Dialogue Agents. Proceedings of the ACL.* *http://www.disc2.dk*
Impact	*Evaluation and usability play a significant role for the technology acceptance through the general public. Usability evaluation standards therefore yields a considerable economic impact.*	
Evaluation		

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Wolfgang Minker; wolfgang.minker@e-technik.uni-ulm.de***
Milestone we asked you to describe	***Gestures and Multimodal Data***
▪ *Creation and Availability of Behavioral Data Resources* ▪ *Uniform Data Annotation Tools and Schemes* ▪ *Common Multimodal Spoken Language Dialogue Systems Development and Evaluation Platforms* ▪ *Usability Evaluation Standards for Multimodal Spoken Language Dialogue Systems*	
Comments	
None.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	Carol Peters	carol.peters@isti.cnr.it
Short name of sub-goal	Cross-Language User Needs Study	
Description of the goal	Despite much work by R&D community on development of CLIR systems, there is surprisingly little take-up so far by the application communities, e.g. so far this technology has not been adopted by any of the large Web search engines and very few commercial information services offer CLIR as a standard service? An extensive study of potential system deployers is needed to identify who are the current and future users of CLIR systems and what are their requirements. The goal to be achieved should be broken down as follows: identification of a set of distinct user group contexts (e.g. intranets of multinational companies; international e-commerce; e-learning, globally distributed digital libraries; tourist information via the web, etc.); for each user group identified, a set of CLIR usability parameters (e.g. efficiency, effectiveness and user satisfaction) should be defined and at least 10 subjects per group should be studied; both questionnaires and hands-on sessions are needed; all aspects of CLIR systems must be covered; interface design and system functionality should be separated and individually surveyed/tested.	http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-12-petrelli.pdf
Expected year of completion	2005	
Justification	The research community has already begun to think about this issue and some initial studies have been made. However, much more needs to be known and well-organised user studies need time to set up.	http://clef.iei.pi.cnr.it:2002/deliv_avail_to_public/Del111.pdf Petrelli, Hansen, Beaulieu & Sanderson. User Requirement Elicitation for CLIR. Proc. ISIC 2002, Lisbon.
Main obstacles for achieving the goal	1. a preliminary investigation would be needed to identify the different user groups that should be involved in such a study to ensure good coverage; 2. user studies are hard as they are	

	time/resource consuming, and difficult to set up/conduct in an objective way; 3. most existing CLIR systems are lab-implemented batch systems, and most operating systems are of limited scope – this limits the setting up of comprehensive hands-on user studies.	
Prerequisites	operational CLIR systems	
Impact	the results of user studies are essential to enable developers to work on bridging current gap between R&D and application world	
Evaluation	*NA*	

our question	*your answer*	*references*
Your name	*Carol Peters*	*carol.peters@isti.cnr.it*
Short name of sub-goal	multilingual text retrieval (MTR)	
Description of the goal	The goal is the development of truly multilingual text retrieval systems, i.e. systems that can query and process collections in multiple languages, rather than simple L1 to L2 querying. The issues involved in L1 to L2 querying have been widely studied and are generally well understood. Truly MTR raises 2 problems which need to be studied in depth: (i) most appropriate system architecture for MTR systems; (ii) translation bottleneck when handling many languages for which language/translation resources (L/TRs) do not exist or are inadequate. Wrt (i) 2 alternatives are currently recognized: queries are processed in 2 steps – translation and retrieval - and separately for each language in target collection, results are then merged BUT no satisfactory merging algorithm has yet been identified; a unified framework can be adopted in which the separate steps are considered as an integrated process and searches are on a single collection containing all languages thus avoiding the merging problems, appropriate modeling tools must be investigated for this purpose, e.g. Bayesian network or language models). The goal should be to conduct a series of comparative studies between these two architectural approaches over a period of 2-3 years, using the same evaluation task as the basis for comparison in order to establish the pros and cons of each approach. Wrt (ii) three paths could be followed to help to overcome the translation bottleneck: development/optimization of methods for creating/improving L/TRs rapidly and cheaply; development/optimization of pivot language methods; development of language independent methods. The TIDES surprise language effort has done much in the first area; a number of groups have already tried the use of pivot languages with varying degrees of success; most of the work on language independent methods so far has been	(i) Nie, J-Y. Towards a Unified Approach to CLIR and Multilingual IR. http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-04-nie.pdf; Nie, J-Y. Query expansion and query translation as logical inference, Journal of the American Society for Information Science and Technology, 54(4): 335-346, 2003. (ii) TIDES Surprise Language Exercise http://language.cnri.reston.va.us/ TeamTIDES/tt02e3-final.pdf P. McNamee, J. Mayfield, and C. Piatko, *`A Language-Independent Approach to European Text Retrieval.* In Carol Peters (ed.), Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000. Ballesteros, L.: Cross-Language Retrieval via Transitive Translation. In Croft. W.B. led.): Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic Publishers, Boston, 2000

	done using n-grams on languages with common origins with considerable focus on named entities. The goal is to understand the issues involved in each of these lines of research and to develop an initial set of guidelines as to how to implement an MTR system when L/TRs are lacking for some of the languages involved.	Lehtokangas, R., Airio, E. Translation via a Pivot Language Challenges Direct Translation in CLIR . http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-07-lehtokangas.pdf
Expected year of completion	*(i) 2007* *(ii) 2008*	
Justification	For (i) 2-3 years should be sufficient to have a clear idea of the pros and cons of the 2 alternate system frameworks For (ii) more time is needed in order to develop and test methodology sufficiently to be able to produce useful guidelines.	
Main obstacles for achieving the goal	The effort involved in the organization of such comparative studies would be considerable and funding would be needed. One ideal platform could be an EC-NSF/DARPA funded collaboration.	
Prerequisites	The proposals above are very high level and involve the development of many tools	
Impact		
Evaluation	The multilingual information retrieval tracks organized by both CLEF and NTCIR could be designed specifically to test the results of the systems/technologies discussed above by offering tasks which involve querying a document collection containing a number of languages and including languages with few L/TRs	http://research.nii.ac.jp/ntcir/workshop/work-en.html www-clef-campaign.org

our question	*your answer*	*references*
Your name	*Carol Peters*	*carol.peters@isti.cnr.it*
Short name of sub-goal	cross-language multimodal systems	
Description of the goal	CLIR must progress from text retrieval to processing queries over languages in multi-media. In general multimedia content is a combination of visual and audio material, either or both which may contain a natural language related component. The non-linguistic elements can be regarded as language independent (ignoring subtleties of cultural interpretation) and one can think of language independent audio and visual search-by-example tools, the language related elements require robust CLIR methodology. This goal could be achieved in two stages. At the end of stage 1, systems would be developed capable of retrieving relevant documents in collections that contain images and/or speech using particular forms of cross-language text retrieval, which works reliably in the face of speech recognition or OCR errors or on short textual captions. The first work of this type has been reported at CLEF2003 for both image and spoken document collections. The target for stage one, would be prototype systems that can accept queries in any of ten different languages (both European and Asian languages) and find relevant documents in English target collections of multimedia documents with 80% of monolingual system performance. At the end of stage 2, systems would be able to combine the results of text-based retrieval with content-based retrieval for image collections, or would be able to take spoken queries as input and use them to search on transcriptions of spoken documents in another language. Testing should be done for target collections in five different languages.	Clough, P., Sanderson, M. The CLEF 2003 Cross Language Image Retrieval Task. http://clef.iei.pi.cnr.it:2002/2003/WN_web/45.pdf Sanderson, M., Clough, P.Eurovision – an image-based CLIR system. http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-14-sanderson.pdf Federico, M., Jones, G. The CLEF 2003 Cross-Language Spoken Document Retrieval Track http://clef.iei.pi.cnr.it:2002/2003/WN_web/ 50.pdf
Expected year of completion	2008 for first results of stage 2	
Justification	There is particular commercial interest in both CLIR image and speech applications. This is	*www.clef-campaign.org*

	perhaps an area where the R&D community has not been meeting the expectations of the application world. This fact should help to encourage fast progress. Also CLEF is putting considerable effort into stimulating advances in this area.	
Main obstacles for achieving the goal	A main difficulty is the acquisition of suitable test collections. For CLIR on image collections, the main obstacle is gaining access to appropriate collections for system development and testing. Unlike, for example, out-of-date newspapers, image collections generally have a strong commercial value and thus it is not easy for the research community to gain access free-of-charge. For CLIR on spoken documents, a major obstacle is the development of good speech processors for many languages rather than just the favoured few. At the moment it is very difficult to find collections of a sufficient size for system development and testing in languages other than English.	
Prerequisites		
Impact	The development of combination systems of the type described above (cross-language retrieval on text AND images AND speech)that involve the interplay of language-dependent and independent factors would be a major step towards the implementation of commercially viable next-generation CLIR systems.	
Evaluation	CLEF should continue to include evaluation tracks for cross-language retrieval on image and spoken document collections, progressively making the tasks more complex and progressing from special types of text retrieval to tasks that involve combining the results of text and image/speech processing and retrieval.	http://ir.shef.ac.uk/imageclef2004/index.html http://hermes.itc.it/clef-sdr04.html

our question	*your answer*	*references*
Your name	*Carol Peters*	*carol.peters@isti.cnr.it*
Short name of sub-goal	multilingual question answering	
Description of the goal	The goal is to develop cross-language systems capable of extracting relevant and precise information from the target collection(s) rather than whole documents. This goal should be achieved in two steps which can, however, be carried out in conjunction, with results for step one providing input for the improvement of results in step 2. The first step involves the development of monolingual QA systems for a number of languages. So far most research in QA has been done on English texts. Procedures that work for English have to be adapted for other languages. The target for this step is prototype monolingual QA systems developed and tested for ten different languages (both European and Asian languages). Step 2 involves the development of prototype cross-language QA systems capable of querying the target collections in the ten languages of step one in at least five languages and with at least 70% of monolingual performance.	Maybury, M.T. Toward a Question Answering Roadmap. http://www.mitre.org/work/tech_papers/tech_papers_02/maybury_toward/maybury_toward_qa.pdf Magnini et al. The Multiple Language Question Answering Track at CLEF 2003. http://clef.iei.pi.cnr.it:2002/2003/WN_web/36.pdf
Expected year of completion	Step 1: 2005 Step 2: 2007	
Justification	CLEF and NTCIR have both stimulated interest in the QA area for languages other than English. This year NTCIR offers monolingual QA for Japanese and CLEF for seven European languages (not including English)and bilingual for 8 languages (also English). Both steps 1 and 2 should thus be achievable within the dates established.	
Main obstacles for achieving the goal	Multilingual QA involves the combination of methodologies and tools from IR and NLP. Getting the two groups to work together is an important	

	challenge in this task.	
Prerequisites	Many tools and technologies are involved.	
Impact		
Evaluation	CLEF and NTCIR should work together in designing evaluation tasks in order to achieve the goals set above.	http://clef-qa.itc.it/2004/ http://www.nlp.is.ritsumei.ac.jp/ qac/index-e.htm

our question	*your answer*	*references*
Your name	*Carol Peters*	*carol.peters@isti.cnr.it*
Short name of the goal	cross-language interactive systems	
Description of the goal	CLIR is not just concerned with system performance judged in terms of the relevance of a ranked list of documents returned in response to a query. The user searching for information in languages with which he has little or no familiarity needs assistance both in formulating and refining his query and in interpreting the results of the search. Thus research is needed into how systems can best help the user in the query formulation and the document selection tasks. The ultimate goal of this task is the implementation of prototype end-to-end multilingual multimedia systems running in real-time which help the user to find relevant information rapidly and interpret it easily. An intermediate goal would be the development of a prototype on-line multilingual text retrieval system searching on document collections in at least five languages with functionality for user-assisted query formulation, refinement, document selection and interpretation.	Oard, D.W., Gonzalo, J. The CLEF 2003 Interactive Track. http://clef.iei.pi.cnr.it:2002/2003/WN_web/31.pdf Gonzalo, J. Scenarios for Interactive Cross-Language Retrieval Systems. http://ucdata.berkeley.edu/sigir-2002/sigir2002CLIR-13-gonzalo.pdf
Expected year of completion	*2008*	
Justification	This is a very hard task.	
Main obstacles for achieving the goal	Studies that involve the user are difficult to organize and resource-consuming. Sufficient funding is needed.	
Prerequisites	Many tools are needed to implement the system, some already exist, others need to be developed: the most ambitious are tools for multilingual multi-document summarisation	
Impact		
Evaluation	An extension of the work done by the interactive track at CLEF with a 4-year program involving tasks of increasing complexity in order to stimulate the development of systems capable of achieving the goal described above.	*http://nlp.uned.es/iCLEF/*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Carol Peters carol.peters@isti.cnr.it***
Milestone we asked you to describe	***cross-lingual information retrieval (henceforth in these templates termed cross-language information retrieval or CLIR)***
cross-language user needs study; multilingual text retrieval; cross-language multimodal systems; multilingual question answering; cross-language interactive systems	
Comments	
The ultimate goal (or grand challenge) for cross-language information retrieval, as first defined at the AAAI-97 Spring Symposium Cross-Language Text and Speech retrieval Workshop, is the development of fully multilingual, multimodal information retrieval systems. Such systems should be capable of processing a query in any medium and any language, finding relevant information from a multilingual multimedia collection, containing documents in any language and form, and presenting it in the style most likely to be useful to the user. Despite the considerable advances, mainly in cross-language text retrieval since then, this goal remains a long-term vision. For the medium term we can envisage the development and testing of the main components of such systems through the fulfillment of a series of subgoals as listed above. It is evident that each of these sub-goals actually represents a main objective in itself and should eventually be structured in a series of subtasks.	

Template 1: description of sub-goals, 1 form for each sub-goal

Our question	*Your answer*	*References*
Your name	*Andrei Popescu-Belis*	*andrei.popescu-belis@issco.unige.ch*
Short name of the goal	*MT Evaluation Framework*	*FEMTI is a first attempt (http://www.issco.unige.ch/projects/isle/femti)*
Description of the goal	*Definition of a coherent framework that groups metrics for machine translation evaluation. The framework consists of weighted links from the various requirements set by an MT user towards the quality metrics that should be used to test whether an MT system fulfills those requirements. The weights, i.e. the relevance of each metric to one or more requirements, must be set by experts of the field and validated by users. The framework could have the aspect of an interactive website that would generate en evaluation plan on user requirements.*	
Expected year of completion	*2007*	
Justification	*The need for such a framework was acknowledged explicitly in (Hovy 1999), and a first attempt, FEMTI, was made during the ISLE project (1999-2002).*	*Cf. URL above.*
Main obstacles for achieving the goal	*The need to poll a significant number of qualified experts. The absence of metrics for some aspects of MT quality. The need to experiment with such a classification in a significant number of case studies.*	
Prerequisites	*MT evaluation metrics for various aspects of quality.*	
Impact	*As the quality of the fully automated tools for MT increases, such a framework will allow for a better tuning of the systems, and possibly for competitive evaluation of heterogeneous systems. The framework will help to organize the market for standalone or embedded MT tools.*	
Evaluation	*The evaluation of such a tool is quite indirect, since it is an evaluation tool. Its frequent use and the satisfaction of the users who made choices based on the framework are two possible indicators of success.*	

Our question	*Your answer*	*References*
Your name	*Andrei Popescu-Belis*	*andrei.popescu-belis@issco.unige.ch*
Short name of the goal	*Automated MT Evaluation Metrics*	*BLEU (Papineni et al. 2001) is a well-known example, used by NIST (USA) in recent MT evaluation campaigns.*
Description of the goal	*Definition of one or more metrics that would automatically assess the "overall quality" of a text translated by an MT system. While quality has several aspects (e.g., syntactic correctness, semantic fidelity, informativeness, etc.), here the goal is to find an automatic metric that would best match the overall judgment of quality expressed as a single rating by human judges (bilinguals judge with the access to the source texts).*	
Expected year of completion	*2006*	
Justification	*Such a metric would allow system developers to test their MT systems often (e.g., daily) for improvements. The need for such a metric (of which some instances are already in use) has become more important as statistical MT systems are used more and more often. The tuning of such systems requires a rapid measure of overall quality rather than a detailed error report that is slower and more expensive to produce.*	
Main obstacles for achieving the goal	*The main problem is of course the absence of a gold standard translation to which a candidate translation could be compared. Therefore, the current attempts use a set of (professional) human translations as a reference, and attempt to compute the distance of a candidate translation to it. The consensus on a given metric can also be an obstacle.* *A more theoretical problem is that is such a metric was easy to compute automatically, it could be used as a learning criterion for statistical systems, therefore helping to solve the problems of machine translation itself.*	
Prerequisites	*While the present attempts are based on lexical (n-gram) distance, more complex automated metrics could require parsers, semantic taggers, and the availability of parallel corpora, or of multiple translation corpora.*	
Impact	*Such a metric would enable developers to test their MT systems quickly and cheaply, which should accelerate the development of high-quality*	

	systems.	
Evaluation	*The evaluation of such an evaluation metric is based on the comparison of its results with human assessment of quality, on a significant corpus of translations graded by humans. A set of coherence criteria for evaluation metrics should be fulfilled too.*	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Andrei Popescu-Belis*** ***andrei.popescu-belis@issco.unige.ch***
Milestone we asked you to describe	***Machine Translation Evaluation***
MT Evaluation Framework *Automated MT Evaluation Metrics*	
Comments	
This technology is important with regard to MT itself, and should not be considered as a fully autonomous research goal in its own, even if it poses a number of important and difficult challenges.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Florian Schiel*	*schiel@phonetik.uni-muenchen.de*
Short name of the goal	*Standardized non-telephone speech corpora*	
Description of the goal	*From our experience producing speech corpora over the last decade we found that non-telephone speech, that is speech recorded in a real-life situation (command&control, communication, data retrieval) are much more difficult to produce than read speech over the phone. Although we have produced many numbers of small, very specialized of such corpora, this is not a very effective way. Better would be a standardized collection of technical settings ('scenarios') and tasks ('domains') that should be covered for each European language in one large controlled speech data collection. Video should be recorded whenever feasible.* *CGN is a good example but it lacks the variety of scenarios. SmartKom was a good example for a number of special scenarios varied in several domains.*	
Expected year of completion	*2010*	
Justification	*Experience from other collections; the total amount of this data collection will probably exceed 50 TB*	
Main obstacles for achieving the goal	*Standardization across all European languages; funding from national sources (since EU cannot be expected to fund such a large enterprise*	
Prerequisites	*UMTS transmitting speech and video*	
Impact	*Speech recognition (command&control, dialogue systems), multimodal Speaker verification*	
Evaluation		*** same*

our question	*your answer*	*references*
Your name	*Florian Schiel*	*schiel@phonetik.uni-muenchen.de*
Short name of the goal	*Standards for Pronunciation coding in SAM-PA for all European languages*	*www.bas.uni-muenchen.de/Bas/BasGermanPronunciation/*
Description of the goal	*Although SAM-PA enables to codify all European (and most other) languages, everybody who is in charge with producing so called 'canonical pronunciation dictionaries' knows that this solves only half of the problem. For every language there are several special rules to observed to yield consistent transcriptions. For German BAS has defined s standard which is now used for all BAS speech corpora and BAS dictionaries. It would be essential that this is done for all European languages.*	
Expected year of completion	*2004*	
Justification	*Mainly intellectual work; no large funding necessary.*	
Main obstacles for achieving the goal	*The main problem is to find an expert for each language who is willing to be responsible for one language and to publish and maintain the standardization.*	
Prerequisites	*Knowledge and expertise*	
Impact	*Speech recognition* *Speech synthesis*	
Evaluation		

our question	*your answer*	*references*
Your name	*Florian Schiel*	*schiel@phonetik.uni-muenchen.de*
Short name of the goal	*Very large pronunciation dictionary*	*www.bas.uni-muenchen.de/Bas/BasPHONOLEXeng.html*
Description of the goal	*Although there exist pronunciation dictionaries for several European languages, these resources are* *- error prone* *- inconsistent with regard to encoding within and between each other* *- not covering more than 30% of day to day language* *- static (on contrast to dynamically updated)* *- in some cases too expensive* *- do not guarantee to cover speech corpora* *Instead of producing a pronunciation dictionary to each new speech corpus it would be much more effective to have a single, constantly maintained resource that covers all resources of one European language. If possible it should be extended by a basic set of the 1 Mio most common used words of that language, all known first and family names, all street/city/state/county/department names of the countries in question. The pronunciation should be marked as being manually produced according to a standardized rule set or being produced automatically (by which software). The resource should be constantly maintained and updated to new words of the language.*	
Expected year of completion	*never (ongoing enterprise)*	
Justification	*-*	
Main obstacles for achieving the goal	*Funding* *Finding institutions for each European language that are capable to maintain such a resource forever and have the expertise to do it.*	
Prerequisites	*Expertise*	
Impact	*Speech recognition* *Speech corpus production* *Speech synthesis*	
Evaluation		

our question	*your answer*	*references*
Your name	*Florian Schiel*	*schiel@phonetik.uni-muenchen.de*
Short name of the goal	*IMDI meta data descriptions*	*www.mpi.nl/IMDI/*
Description of the goal	*One of the greatest problems still is to find an existing speech resource. A distributed meta data descriptor system like IMDI would make this problem smaller. Therefore all European language resources should have at least a minimum descriptor in the IMDI hierarchy.*	
Expected year of completion	*never (ongoing enterprise)*	
Justification		
Main obstacles for achieving the goal	*How to force producers and maintainers to provide the IMDI files* *Funding is NOT the problem here.*	
Prerequisites	*IMDI Tools of MPI Nijmegen*	*www.mpi.nl/IMDI/tools/*
Impact	*Speech technology in general*	
Evaluation		

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Florian Schiel*** ***schiel@phonetik.uni-muenchen.de***
Milestone we asked you to describe	*speech resources*
- Standardized non-telephone speech corpora for all European languages *- Standards for Pronunciation coding in SAM-PA for all European languages* *- Very large and dynamically updated, standardized pronunciation dictionaries for all European languages* *- IMDI meta data descriptions to all existing speech resources*	
Comments	
*** whatever comments you have*	

Template 1: description of sub-goals, 1 form for each sub-goal

Our question	*Your answer*	*References*
Your name	*Marc Schröder*	*schroed@dfki.de*
Short name of the goal	*Emotional speech databases*	Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication Special Issue Speech and Emotion, 40(1-2):33-60.
Description of the goal	*Databases of spontaneous emotional speech, representative for typical application scenarios, annotated using emotion representations suitable for emotion recognition and generation tasks*	
Expected year of completion	*2007*	
Justification	*Current efforts under way in HUMAINE WP5*	*http://emotion-research.net*
Main obstacles for achieving the goal	*Ethical and methodological difficulties of obtaining spontaneous emotional data; a sufficiently large-scale database would require considerable funding; copyright issues*	
Prerequisites	*Suitable emotion representations; ethical guidelines on data collection; recording and labelling paradigms*	
Impact	*Enables data-based determination of emotional features for emotion recognition and generation*	
Evaluation	*How spontaneous and natural? How many speakers? How many emotions per speaker? How "objectively" are emotions annotated?*	

Our question	*Your answer*	*References*
Your name	*Marc Schröder*	*schroed@dfki.de*
Short name of the goal	*Emotion recognition*	*ERMIS: http://www.image.ntua.gr/ermis; HUMAINE WP4: http://emotion-research.net*
Description of the goal	*Recognition of emotions and emotion-related states (e.g., arousal) from speech and from text.*	
Expected year of completion	*2009*	
Justification	*Data should be available until then; conceptual issues such as emotion representation should also be sorted out sufficiently by then*	
Main obstacles for achieving the goal	*Acoustic similarity of very different emotions (e.g., anger/joy); emotion representations must be used which capture these effects*	
Prerequisites	*Statistical methods for automatic classification; suitable acoustic parameter sets, and automatic methods for measuring them; enough and good data for training the methods*	
Impact	*Enables emotion-sensitive devices, emotion-detection in security-related environments, emotional human-machine interaction*	
Evaluation	*Can systems deal with expected input? How meaningful is systems' output with unexpected input? A flexible measure of success, taking into account "degree of correctness", would also be required.*	

Our question	*Your answer*	*References*
Your name	*Marc Schröder*	*schroed@dfki.de*
Short name of the goal	*Emotion generation*	*NECA: http://www.ai.univie.ac.at/NECA; HUMAINE WP6: http://emotion-research.net;* Schröder, M. (2001). Emotional speech synthesis: A review. In Proceedings of Eurospeech 2001, volume 1, pages 561-564, Aalborg, Denmark. http://www.dfki.de/~schroed
Description of the goal	*Generation of emotional speech and text.*	
Expected year of completion	*2009*	
Justification	*Emotional data should be available until then; representation issues can be expected to be sorted out by then; it can be hoped that natural and parametrisable speech synthesis technologies are available by then.*	
Main obstacles for achieving the goal	*Acoustic speech synthesis algorithms that are both flexible/parametrisable and natural.*	
Prerequisites	*Suitable emotion representations; speech synthesis technology allowing for acoustic modifications including voice quality while preserving naturalness*	
Impact	*Emotional human-machine interaction; emotionally appropriate announcement systems; believable ECA (embodied conversational agent) systems.*	
Evaluation	*Perception tests, using preference tasks in which several acoustic realisations are combined with emotional text.*	M. Schröder (to appear). Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. *PhD thesis, Institute of Phonetics, Saarland University.* M. Schröder (to appear). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. Accepted for publication in Workshop on Affective Dialogue Systems (ADS 04), Kloster Irsee, June 2004.

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Marc Schröder, schroed@dfki.de***
Milestone we asked you to describe	***Emotions***
Emotional speech databases *Emotion recognition* *Emotion generation*	
Comments	
There is only a partial overlap between the fields of emotion research for human-machine interaction and language technology. The sub-goals listed above are the ones with a strong "language" focus; others, such as "emotion representations", "emotion models" or "emotional interaction", are only cross-referenced to. The network of excellence HUMAINE, http://emotion-research.net , addresses these issues more fully.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Kiril Simov*	*Kivs@bultreebank.org*
Short name of the goal	*Matrix Multilingual Treebank*	
Description of the goal	*The aim is the creation of a set of sentences in several languages annotated with respect to several annotation schemes. The annotation schemes have to cover the main linguistic theories like HPSG, Dependency Grammars, GB, LFG, Construction Grammar, etc. Also the sentences have to be annotated with meta-information about the linguistic phenomena they highlight. Transformation rules between the annotation schemes are desirable. A set of tools and language resources which can support the creation of a treebank for a new language on the basis of the matrix treebank is necessary. Support for evolution of the matrix treebank and the treebanks created on the basis of it is also required.*	*lingo.stanford.edu/matrix/ ; http://www2.parc.com/istl/groups/nltt/pargram/ ; Nivre, J. (2003) Theory-Supporting Treebanks. http://www.msi.vxu.se/~nivre/papers/support.pdf ; Kiril Simov. HPSG-Based Annotation Scheme for Corpora Development and Parsing Evaluation. http://www.bultreebank.org/papers/p60-simov-ranlp03.pdf; and others*
Expected year of completion	*2007*	
Justification	*There are a lot of treebanks created for languages from different language groups and with respect to different annotation schemes. Also there are a number of tools for the creation of treebanks. There are experiments on mapping of different annotation schemes. Some parallel (or comparable) treebanks already exist.*	*www.cis.upenn.edu/~treebank/home.html quest.ms.mff.cuni.cz/pdt/ www.bultreebank.org odur.let.rug.nl/~vannoord/trees treebank.linguist.jussieu.fr/ redwoods.stanford.edu/ www.sfs.nphil.uni-tuebingen.de/de_tuebadz.shtml*

		www.ii.metu.edu.tr/~corpus/treebank.html *www.di.unito.it/~tutreeb/* *www.ims.uni-stuttgart.de/projekte/TIGER/*
Main obstacles for achieving the goal	*A widely accepted linguistic ontology (standardization of the linguistic concepts). Missing tools for mapping of linguistic analyses between different theories.*	
Prerequisites	*Existing treebanks and annotation schemes; mechanisms for reduced annotation effort; off-line transformation of linguistic knowledge; basic language resources and methodology for their implementation*	
Impact	*The matrix multilingual treebank will ensure a cheaper creation of treebanks for languages that lack them. The common model will also facilitate the usage of the constructed treebanks.*	
Evaluation	*The evaluation of the results can be done on the basis of simultaneously comparison with the annotation schemes for already existing treebanks and annotation of new sentences. In this way the coverage with respect to the linguistic phenomena will be controlled.*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Kiril Simov*	*Kivs@bultreebank.org*
Short name of the goal	*Pragmatically Annotated Treebanks*	
Description of the goal	*(A) treebank(s) for one or several languages which contain(s) annotation of the three levels: syntactic, semantic and pragmatic in a common, modular annotation scheme. Besides the syntactic information at least the following information is necessary to be presented: description of the referents (objects and events in the world) including the obligatory implied ones, co-reference relations, ontological classes of the referents, ontological relations between referents, lexical chains, cohesion relations. Special attention will be paid to the intersentential relations.*	*http://www.cis.upenn.edu/~ace/* *http://quest.ms.mff.cuni.cz/pdt/* *www.bultreebank.org* *www.icsi.berkeley.edu/~framenet/* *www.coli.uni-sb.de/lexicon/index.phtml* *Kerstin Kunz and Silva Hansen-Schirra Coreference Annotation of the TIGER Treebank: www.masda.vxu.se/~rics/TLT2003/doc/kunz_hansen.pdf*
Expected year of completion	*2008*	
Justification	*There exist annotation schemes that already incorporate partially the required information.*	*See the above URLs*
Main obstacles for achieving the goal	*Appropriate lexical resources interconnected with ontologies are still underdeveloped.*	
Prerequisites	*Existing treebanks; domain based ontological lexicons; generic schemes for semantic annotation of text; top-level ontologies; annotation of pragmatic content; approaches for markup of discourse structure and pragmatics; superficial semantic processing based on ontological lexicons*	
Impact	*Such treebanks will facilitate the development of robust deep analysis for tasks such as: Information management (retrieval, extraction, summarization), question answering*	
Evaluation	*The evaluation will follow the standard measurement inter-annotators agreement and adaptation of the methods for evaluation developed for the syntactically annotated corpora.*	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Kiril Simov, kivs@bultreebank.org***
Milestone we asked you to describe	Treebanks
1. Matrix Multilingual Treebank *2. Pragmatically Annotated Treebank (reference, lexical chains, ontological relations, cohesion relations)*	
Comments	
I divided the goal into two sub-goals: *The first one is oriented towards unification of the existing approaches to treebank creation in order to minimize two things mainly: (1) the creation of a treebank for a new language with minimal effort what is especially important for less-spoken languages; and (2) to improve the usability of the treebanks.* *The second sub-goal is towards the extension of the linguistic knowledge encoded in the treebanks.*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Tokunaga, Takenobu*	*take@cl.cs.titech.ac.jp*
Short name of the goal	*Multilingual Lexicon*	*** URL or publication (could be one of your own) supporting or clarifying your point*
Description of the goal	*A multilingual lexicon of 200,000 entries for the 20 main languages including Asian languages that could be usable for machine translation systems working on personal computers.*	*** same as above*
Expected year of completion	*2015*	*** same*
Justification	*• 3 years for defining the specification of multilingual entries,* *• 2 years for building basic 5,000 entries, including revision of the entry specification* *• 5 years for scaling up to 200,000 entries*	*** same*
Main obstacles for achieving the goal	*• Defining the specification of lexicon would be the biggest obstacle.* *• In choosing languages, various factors should be taken into account, such as political issues, market size, research level and so on.* *• Scaling up requires enormous amount of money.*	*** same*
Prerequisites	*There have already been such efforts such as EAGLES and ISLE/MILE proposals. However their main target is European languages. We tried to apply such proposals to several Asian languages and found some irrelevancy. These proposal would be a good starting point to defining the specification of lexicon.*	*See ISLE/MILE final workshop discussion.*
Impact	*Such kind of language resources would impact on multilingual machine translation systems.*	*** same*
Evaluation	*• Quantitative evaluation: the number of entries and languages* *• Qualitative evaluation: translation quality when used in translation systems*	*** same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	*Tokunaga, Takenobu (take@cl.cs.titech.ac.jp)*
Milestone we asked you to describe	*Multilingual lexicon*
• Defining the specification of multilingual entries, *• Building basic 5,000 entries, including revision of the entry specification* *• Scaling up to 200,000 entries*	
Comments	
In parallel with building a multilingual lexicon, it would be interesting to build multilingual phrase book (translation memory) which is usable for human translator. To achieve this goal, it is necessary to realize a framework supporting distributive information entry with maintaining its consistency and quality.	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Nicoletta Calzolari*	*glottolo@ilc.cnr.it*
Short name of the goal	*Computational Lexicons – Open and Distributed Lexical Infrastructure*	
Description of the goal	*Definition and creation of an Open and Distributed Lexical Infrastructure on the web, where lexical resources are accessible through web services.* This infrastructure will be based on open content interoperability standards, and is seen as the cooperative effort of different types of communities (such as commercial content producers, lexicon producers and users, etc.). It is intended to cover a very large number of European and non-European languages.	
Expected year of completion	*2008*	
Justification	*This is seen as the only way to overcome the problem of broad availability of lexical resources, and as a way to allow integration of lexical resources.*	
Main obstacles for achieving the goal	*The technology is there.* *The willingness of many groups world-wide is there.* *Mainly there are organizational and financial issues, i.e. a cooperative initiative should be financed to make this possible.*	
Prerequisites	*Availability of standards (at many levels), their extension and integration when needed.* *Design of a new model of lexical architecture.*	
Impact	*Impact on all HLT where computational lexicons are needed.* *Also a change in the way resources are distributed and commercialized, mainly as a service. Access and pricing policies must be carefully designed.*	
Evaluation	*It is important to have validation protocols for the resources which are part of the infrastructure, e.g. a mechanism of certificates of validity can be designed.*	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	*Nicoletta Calzolari*	*glottolo@ilc.cnr.it*
Short name of the goal	*Computational Lexicons – Dynamic Lexicons: New types of resources which are Corpus and Lexicon together*	*Nicoletta Calzolari, Computational Lexicons and Corpora: Complementary Components in Human Language Technology. In International Congress of Linguists, Prague, 2003.*
Description of the goal	A change of perspective on lexicons as static resources towards dynamic entities, whose content is co-determined by automatically acquired linguistic information from text corpora and from the web. *The acquisition tools must be able to increase the repository with new words/terms, possibly their definitions, domain, sense-in-context, multi-words, etc., from digital material, to learn concepts from text – including automatic multi-lingual thesaurus building, and to tailor resources to specific needs.* *Agents will look for examples, identify uses in monolingual/multilingual web texts for glossary creation.* *This will ensure also virtual links between lexicons and examples: corpus/web samples, image samples, clips and videos, etc., and will allow the creation of a new generation of "lexicon-corpus resources" together.*	
Expected year of completion	*2008 (for a good prototype)*	
Justification	*No static lexicon can ever be 'complete', for theoretical reasons. Static core lexicons must be enriched, tuned, etc. with lexical information automatically acquired and customized to different domains/applications, etc., otherwise coverage and/or accuracy will remain inadequate.*	
Main obstacles for achieving	*This implies focused involvement of research groups in the machine learning community, developing new and*	

the goal	*strong algorithmic methodologies to model textual statistics, and integrating them with traditional NLP tools.*	
Prerequisites	*Robust (semi)-automatic or machine aided methods must be used wherever possible in resource work. The* increasing availability and reliability of robust techniques (for chunking, shallow parsing, functional analysis, named entity recognition, etc.), and the ability to integrate them, makes the exploitation of text corpora of greater relevance in many HLT tasks, and allows the acquisition of lexical information which complements that available in static lexicons. *A lexicon model which is suitable to accomodate the information automatically acquired.*	
Impact	*Impact on all HLT where computational lexicons are needed.*	
Evaluation	*It is important to have validation protocols for the acquired resources.*	

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	
Milestone we asked you to describe	***Computational Lexicons***
Computational Lexicons – Open and Distributed Lexical Infrastructure	
Computational Lexicons – Dynamic Lexicons: New types of resources which are Corpus and Lexicon together	
Comments	

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	Stelios Piperidis	spip@ilsp.gr
Short name of the goal	Parallel Corpora and multi-level alignment	Parallel Text Processing, Alignment and use of translation corpora, Veronis, J. (Ed), Kluwer Academic Publishers, Text Speech and Language Technology Series
Description of the goal	Parallel corpora of ca 10M words in 10 different domains, for the main language pairs and of such quality that they can be used for multilingual resources elicitation purposes (glossaries, lexical and grammars) and machine translation purposes	** *same as above*
Expected year of completion	2006-2010	** *same*
Justification	• implementation of the EU public sector information directive • increased demand for multilingual applications	** *same*
Main obstacles for achieving the goal	• legal and copyright issues • varying degree of parallelness of Web documents • sparseness of useful data for interesting applications • such corpora are mainly available through international organizations resulting in distortions in language use	** *same*
Prerequisites	• text alignment tools • existing glossaries/lexica to bootstrap the word alignment process • pos tagging and possibly chunking tools for word and phrase alignment • statistical models	** *same*
Impact	• All multilingual applications • Automatic corpus-based glossary/lexicon building • Transfer grammar induction • Cross-lingual information retrieval	** *same*

	• Machine Translation (both statistical and rule-based) • Computer-assisted language learning (CALL)	
Evaluation	Use of reference data to enable computing information retrieval driven measures: precision, recall, F-measure	** *same*

Template 1: description of sub-goals, 1 form for each sub-goal

our question	*your answer*	*references*
Your name	Stelios Piperidis	spip@ilsp.gr
Short name of the goal	Comparable Corpora and word alignment	Parallel Text Processing, Alignment and use of translation corpora, Veronis, J. (Ed), Kluwer Academic Publishers, Text Speech and Language Technology Series
Description of the goal	Comparable corpora of ca 30M words in different domains, for the main language pairs and of such quality that they can be used for bilingual glossary/lexicon resources elicitation purposes	** *same as above*
Expected year of completion	2006-2010	** *same*
Justification	• implementation of the EU public sector information directive • increased demand for multilingual applications	** *same*
Main obstacles for achieving the goal	• legal and copyright issues	** *same*
Prerequisites	• existing glossaries/lexica to bootstrap the word alignment process • pos tagging and possibly chunking tools for word alignment • statistical models	** *same*
Impact	• All multilingual applications • Automatic corpus-based glossary/lexicon building • Cross-lingual information retrieval	** *same*
Evaluation	Use of reference data to enable computing information retrieval driven measures: precision, recall, F-measure	** *same*

Template 2: summary list of sub-goals

Your decomposition of the goal into sub-goals	
Your name and email	***Stelios Piperidis, spip@ilsp.gr***
Milestone we asked you to describe	***Parallel Corpora***
Parallel Corpora and multi-level alignment Comparable Corpora and word alignment	
Comments	
The usefulness of Parallel Corpora and their processing lies on their high multiplier effect as derivative resource generators. The challenges with parallel corpora are both the sparseness of useful and interesting data, if one excludes institutional texts (e.g. EU texts), and the relative difficulty in building tools that generate useful derivative resources with high accuracy. *In the subgoals above corpora are intertwined with the associated tools.*	