-Technology Survey-
Multilingual and Multimodal Spoken Dialogue Systems

In what follows is a review of the EPOCH relevant research areas, techniques and technologies involved in Natural Language Processing (NLP) for the design of Human-Computer Interactive (HCI) systems for Cultural Heritage (CH).

Speech Recognition

If spoken input is to be used as one modality, we must employ a Speech Recognizer (SR). Some aspects of what the SR delivers have a direct impact on the subsequent modules in the system. SR may be speaker dependent; in this case a training phase is required for the SR to get used to the speaker voice. Speaker independent SRs supporting wide varieties of languages are therefore preferable for Cultural Heritage applications. One key factor in choosing a SR is the Word Error Rate (WER), which is the percentage of words which are typically misrecognized by the SR, which is around 30% for state of the art SRs. Because CH applications are mostly domain-oriented, as opposed to task-oriented, it is also highly desirable that the SR allows the setting of recognition resource parameters (such as grammars) dynamically during run-time.

Typically, the format of the output of the SR is made up of words, out of vocabulary and pause symbols (OOV, PAUSE). It may also contain capital letters, common expressions combined as a single word ("Ducal_Palace"), short forms ("I'm"), hyphenated words ("short-sighted") and words between quotation mark (""). The SR produces a typical N-best list or Word Graph (WG) with confidence scores for words and for the whole sentence. Each path in a word graph represents an alternative sentence that the user could have said. Each word in a path has a confidence score, and each path is given an overall confidence score.

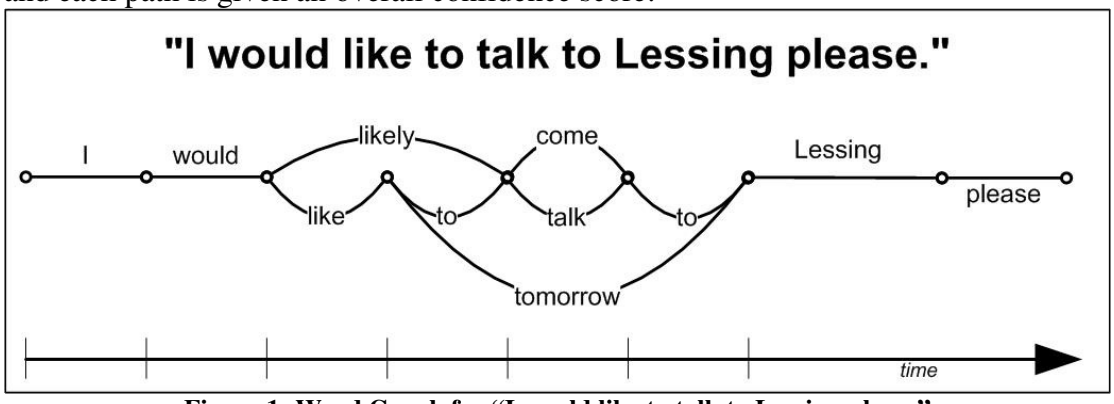

Figure 1: Word Graph for "I would like to talk to Lessing please"

Referring to Figure 1, the following five paths result from the WG:

1. I would like to talk to Lessing please.
2. I would like to come to Lessing please.
3. I would like tomorrow Lessing please.
4. I would likely come to Lessing please.
5. I would likely talk to Lessing please.

Semantic Parsing

The list of words coming out of the SR is to be interpreted semantically. The problem of finding a good interpretation of natural language utterances is particularly difficult to solve. *Semantic parsing*, as it is known, is described in [Allen, 1995] as the process of mapping a natural language input (a sentence) to some structured meaning representation which is suitable for manipulation by a machine. In general, Natural Language Interfaces are difficult to build and must be tailored to each domain of application. Semantic parsing is challenging because it involves the concept of *Natural Language Understanding (NLU)*. Classical methods use hand-written rules and formal semantics to build up a suitable representation. Besides the heavy burden of finding appropriate rules for this task, formal semantics relies on the principle of *compositionality*, which does not take into account the non negligible idiomatic nature of natural language, especially in conversational speech. Not so long ago, corpus-based methods [Ng and Zelle, 1997] received much attention for addressing these problems. They have been applied with success in areas like speech recognition [Rabiner, 1989, Bahl et al., 1983], part-of-speech tagging [Charniak et al., 1993], text or discourse segmentation [Litman, 1996] and syntactic parsing [Collins, 1997, Manning and Carpenter, 2000]. They allow the construction of systems which satisfy desirable properties of NLP applications. These properties can be summarized as follow [Armstrong-Warwick, 1993]:

- Acquisition, i.e. automatically acquiring knowledge (domain specific or not) that would be necessary for the task
- Coverage, i.e. handling the potentially wide range of possibilities that could arise in the application
- Robustness, i.e. accommodating real data which may not be "perfect" (the noise factor) and still being able to perform reasonably well
- Portability, i.e. easily applicable to a different task in a new domain

Corpus-based methods rely on a training corpus which is, in general, a collection of sentence-meaning pairs (for the task of semantic parsing). Two main areas of corpus-based approaches for semantic parsing can be distinguished: the *Machine Learning* approach and the *Statistical* approach. In the machine learning approach, some learning algorithm is used to learn how to go from utterances to meanings. Statistical approaches collect a number of parameters from the training corpus to help a semantic parser discriminate between good and not so good meanings. Some approaches use both. Traditionally, methods used to construct semantic parsers very often involve the creation of rules by a knowledge expert. These hand-crafted rules make the parser incomplete, even for a specific domain. As knowledge to encode grew in size, handcrafted work became more and more difficult as we approach the so-called *knowledge engineering bottleneck*. The result was a very inefficient and fragile parser. More recent approaches tend to avoid this knowledge-engineering perspective in favour of an empirical, corpus-based approach where parsers are constructed through learning.

Dialogue Management (DM)

With the advent of domain-oriented Spoken Dialogue Systems (SDSs), the task of managing a conversation will probably undergo fundamental change in the next few

years. It is one thing to convey a proper conversation in a task-oriented system, e.g. to help a customer to buy plane tickets or get information on current events, it is quite another to enter in a conversation about the effects of global warming or life in general. By constraining the input, task-oriented systems have achieved a satisfactorily level of functionality. Domain-oriented systems are in their infancy. Both types of conversations may be valuable for cultural heritage applications. A request about items being on display in a museum is a simple task, while discussing with a virtual historical character about its work is a domain-oriented dialogue. One important lesson for CH applications can be drawn from work by [Core et al., 2003] with regards to initiative in tutorial dialogues, where a proposed set of speech acts helps tracking down who has the *initiative.* In [Whittaker and Stenton, 1988], *initiative* is defined as the control taking of the dialogue by one participant. More to the point, [Chu-Carroll and Brown, 1998] differentiate dialogue initiative from task initiative. Dialogue initiative "tracks the lead in determining the current discourse focus" (p.6) and task initiative "tracks the lead in the development of the agent's plan" (p.6). Roughly speaking, besides the purely semantic content of dialogues in conversations, an important aspect of success in speech, measured by the user's satisfaction with the quality of the output, variation, humour, breath and depth of topics, comes from knowing who has the initiative and who should now have it. The main reason appears clear by looking at findings from [Core et al., 2003] which studied initiative management in two dialogue strategies: *didactic* tutoring and *socratic* tutoring. Socratic tutoring "is characterized by the use of questions and other hints to draw out answers from students having difficulty" (p.1), while in didactic tutoring, "the tutor points out the student's error and explains how to derive the correct answer" (p.2). What these findings show is that if we want the student to achieve as much learning as possible in a tutorial dialogue, then it is wiser to adopt a socratic way of carrying out the dialogue by asking more questions, with roughly one out of ten turns being led by the student. As a matter of fact, this will increase student verbosity, and make the dialogue more interactive. Finally, [Core et al., 2003] notes a positive correlation between interactivity and learning. In spontaneous conversation drawn from a large domain, the tutor is replaced by a machine and the goal of the machine dialogue management unit is more blurred (entertainment, education, information, etc.) where learning *per se* is not the final goal, but part of it. Because *verbosity*, *interactivity* and *learning* all contribute to make a dialogue successful, one can adopt a dialogue management policy inspired by [Core et al., 2003]'s findings for tutorial dialogues.

Speech Generation

Most work in Natural Language Generation has focused on written language, as opposed to spoken output [Biber et al., 1999]. The study of speech generation has yet to achieve the same level of maturity of its close relative, and most of the spoken systems have tailored output originally intended for document creation, in other words written language, to be passed along to a plain text-to-speech synthesizer. Tailoring involved mainly that the text to be synthesized is augmented with basic sentence level prosodic information. However, there is more to this than meets the eye; to convey properly the intended semantics as well as the emotion of the speaker and its particular style, we need more than the bare words; *concept-to-speech* systems, as they have been termed, attempt to act directly on the final phonetic representation

based on contextual, linguistic and semantic information to convey semantic, affective and stylistic information. Variable output can be achieved via stochastic generation.

Stochastic Generation

To produce human-like spoken output, one must take into account variability, and this is where language modelling comes into play. Language modelling using trigrams has been around for some time [Bahl et al., 1983] as well as tools for creating the models [Clarkson and Rosenfeld, 1997]. In recognition, language models are used to extract the most probable sequence of words given the acoustic evidence. In generation, the "evidence" we have is the utterance's class, a combination of topic and speech act. We are not looking for the most likely sequence, but rather for the generation of outputs according to the distribution of each gram in the corpus for that particular class. Pioneering work in this area is [Rudnicky and Oh, 2000]. The general idea of the generation algorithm is that words are generated randomly until the end of the sentence marker "</utt>" is generated. It should be clear that whenever a higher order n-gram model is used in generation, the fertility decreases while precision increases; using a lower order model has the exact opposite effect on precision and fertility. The arrows in Figure 2 illustrate the effect of lowering or raising the model's order on precision and fertility.

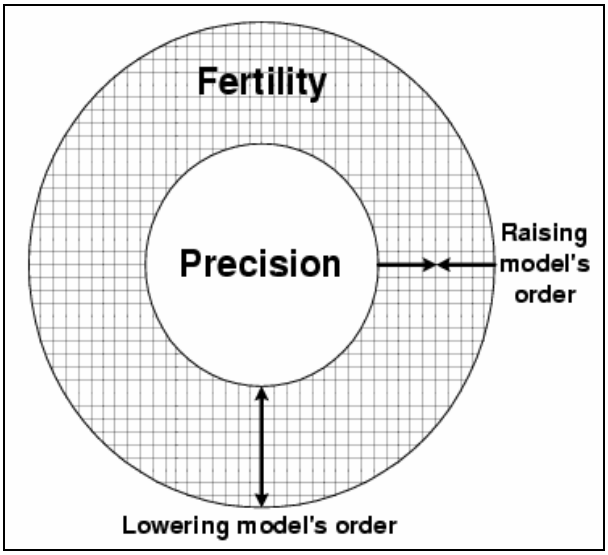

Figure 2: Precision versus Fertility

Multilinguality

Globalization means that interaction will be carried out in different languages.
The main problem with applications that interact in more than one languages is that *interpretation* must be preserved, not (literal) *meaning*. Consider the following German example, uttered by a virtual guide to seek confirmation from a user navigating and interacting through a virtual replica of a historical town: "Geht es bei Dir?". Depending on the target language, the literal meaning rendered in other languages is more or less close to the intended interpretation. In French, automatic translation produced by a machine could give something like « *Va-t-il avec toi* ? » whilst *"Does it go with you?"* could be a possible English translation. French users would probably infer quickly the right interpretation "Ça va?", but the English speakers would probably struggle to interpret the guide's reply in the right way "*Is that alright*

with you?". The difficulty is that interpretation depends on context, but the good news is that interactive systems are considering more and more modalities, as we will see in the next section, which provide a richer source of contexts, for input interpretation and for output generation.

Multimodality

Modern generation systems for speech are set in a *multimodal* environment. In this type of environment, speech is not the only mean of interaction; point and click operations, for example, can be used along with speech. Because each modality interacts with, complements and to some extent modifies one another, modern speech generation must be studied in a multimodal setting. A long-term objective for dialogue systems is "sensory realism", so that humans interacting with a machine could use their senses in a natural way. On one hand, visual and auditory are two modes of feedback in HCI that have been used extensively. On the input side, gesture and speech recognition have progressed significantly over the last few years. However, the touch, taste and smell modalities are yet to be used broadly in HCI. On the other hand, users do not employ modalities in the same way: for example, in the domain of searching newspaper texts, experiments [Klein et al., 2001] show that non-expert users benefit from (time wise) and prefer to combine speech and click operations over written and click operations.

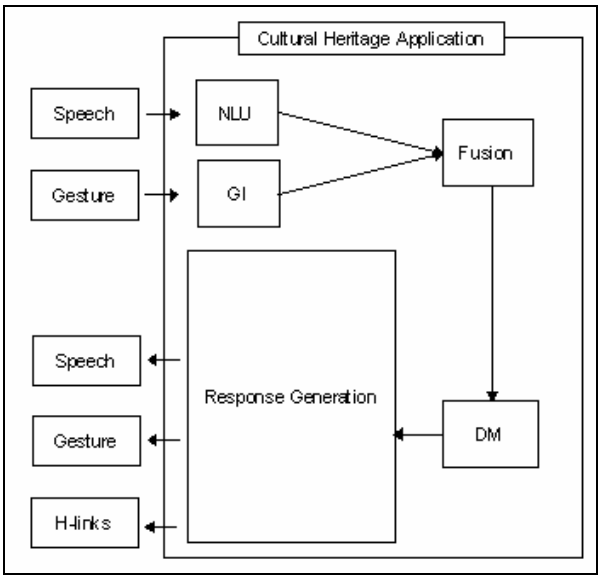

Figure 3: Integrating Modalities (GI = Gesture Interpretation)

Therefore, a major challenge for multimodal applications is the integration of those modalities. To be interpreted correctly and efficiently, parallel modes of expressions need to be integrated in a general and principled framework so that synchronization and further development fall in naturally in the unified model. A typical integration scheme for CH application may look like Figure 3.

Existing relevant projects for Spoken Dialogue Systems in Cultural Heritage

Three applications can provide a very useful framework for the development of Dialogue systems in the Cultural Heritage domain. [WYSIWYM] aims to allow domain experts to encode their knowledge directly. The [NECA] system, originally

designed for combining situation-based generation of natural language and speech, gesture and emotions situated in social communication, provides a general model for the integration modalities and emotion during the integration. The [M-PIRO] project provides a direct example of how a system that generates descriptions of museum objects can be tailored to the user in terms of level of expertise, modality expressed, wording, phrasal complexity and language.

Putting it all together

In Figure 4, we show the data flow for a typical spoken interaction between a user saying "Do you like Harry Potter?" to a virtual tour guide, for which the guide strategy is simply to refocus the attention to a topic he is more familiar with: "Actually, I like von Goethe".

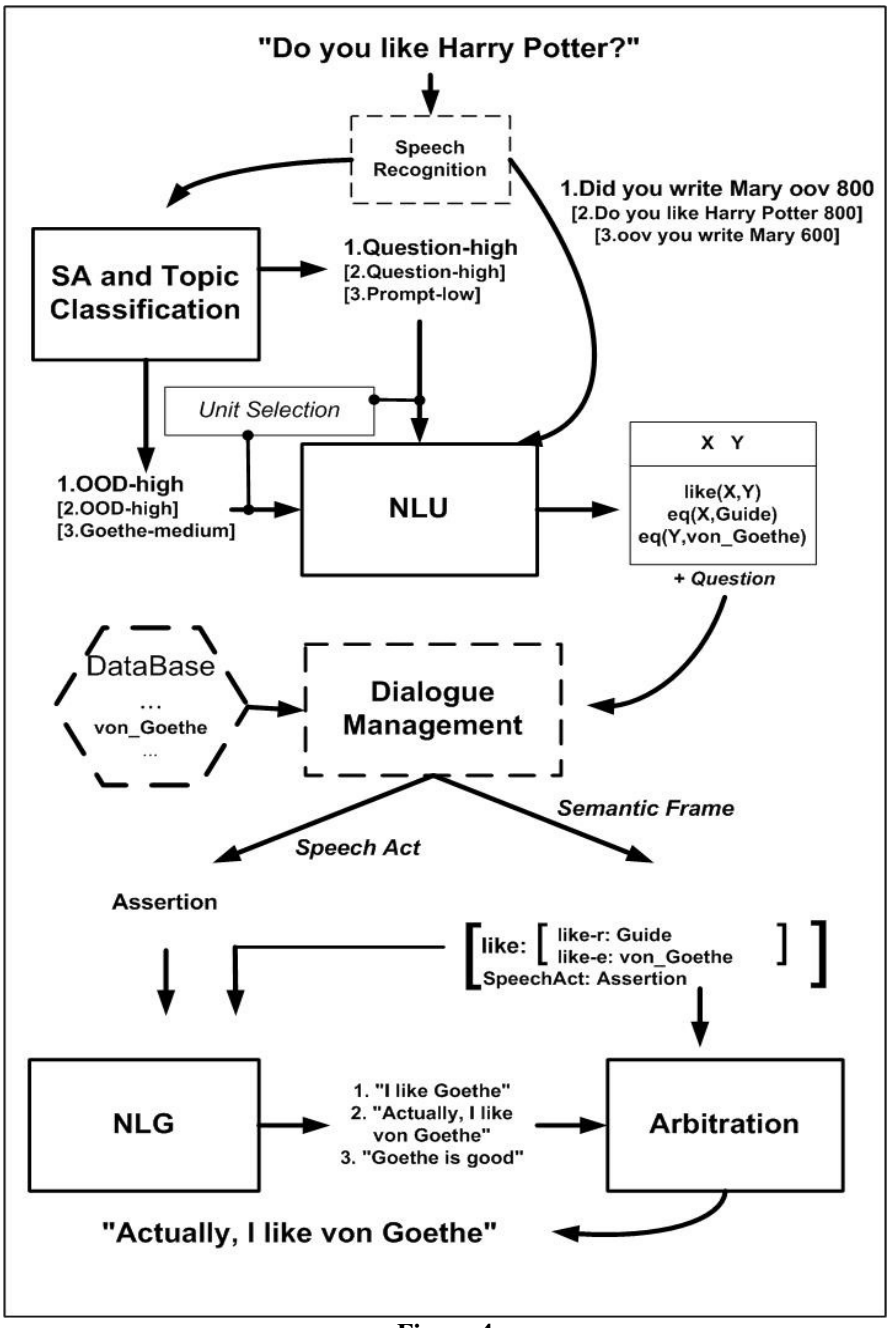

Figure 4

State of the Art: Natural Language Generation in guided tour systems

AT&T NJFun

"NJFUN is a spoken dialogue agent that allows users to access information about things to do in New Jersey, via a telephone conversation". It is not multilingual and does not involve any Virtual Reality (VR). "NJFUN has served as a test bed for demonstrating the use of Reinforcement Learning for optimizing spoken dialogue management. It uses a speech recognizer with stochastic language models trained from example user utterances, and a TTS system based on concatenative diphone synthesis. Its database is populated from a public webpage to contain information about activities".

Reference: Satinder Singh, Diane Litman, Michael Kearns and Marilyn Walker. Optimizing Dialogue Management with Reinforcement Learning: Experiments with the NJFun System. Journal of Artificial Intelligence Research(JAIR), 2002.

IRST AL Fresco

The IRST AL Fresco system is a museum guide that sits on a hand-held device. It takes visitors round an Italian museum, and is multilingual, but with no VR. "ALFRESCO is a prototype with the purpose to exploit the potentiality of integrating natural language with other communicative modalities", so it is multimodal. "The approach relies on the technological possibility of moving in a rich information space with rapid interleaving of different information and different media, and on the high-level integration into a coherent information seeking dialogue".

Website: http://tcc.itc.it/history/projects/alfresco/alfresco.html

ILEX

The ILEX system generates dynamic labels for visitors to a Scottish museum. It is a web based system, is not multilingual and does not involve any VR. The focus of the project is automatic text generation, in order to "produce descriptive, explanatory, or argumentative texts to accomplish various different communicative tasks". The very dynamic nature of the labels "has a number of advantages, such as taking into account the visitor's level of expertise about the objects, as well as the discourse history---the objects which the visitor has already seen---so that information the visitor has already assimilated can be taken into account description of the object currently being viewed can make use of comparisons and contrasts to previously-viewed objects, while omitting any background information that the visitor has already been told".

Website: http://www.hcrc.ed.ac.uk/ilex/

M-PIRO

M-PIRO is an extension of the ILEX project. While "ILEX served up personalized information objects", M-PIRO main advance is that "it develops an authoring tool (developed at NCSR "Demokritos") to help museum creators create and edit the domain-specific knowledge base and linguistic resources". It is multilingual, uses a

museum guide, but no VR, although one partner ("Foundation of Hellenic World") "have been working on a new M-PIRO prototype that will embed the project's technology in their immersive VR system".

Website: http://www.ltg.ed.ac.uk/mpiro/

ICT

On the other side of the Atlantic, one leader in research on speaking avatars is ICT in Marina del Rey, California. A lot of their applications are military and most (if not all) are unilingual, but there is a system in development called "Integrating Architecture" that "will provide an integrated infrastructure for fundamental research in the disparate areas of artificial intelligence (AI), graphics, sound, animation, and immersive display technologies, such as FlatWorld (the Mixed Reality Simulation Space) and the Virtual Reality Theater", which could be relevant for EPOCH.

Website: http://www.ict.usc.edu/

NICE

The NICE project has just completed and a working system can be used by visitors in the Hans Christian Museum in Odense Denmark. "NICE aims to demonstrate universal natural interactive access, in particular for children and adolescents, by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters. The communication consists of domain-oriented spoken conversation combined with 2D input gesture into a 3D dynamic graphics virtual world inhabited by the fairy-tale author Hans Christian Andersen and animated characters from his fairy-tale universe. For the first time, professional computer games technologies are joined with advanced spoken interaction, and speech recognition technology is specially developed for recognising the speech and spoken linguistic behaviour of children and adolescents".

Website: http://www.niceproject.com/

NECA

"NECA promotes the concept of multi-modal communication with animated synthetic personalities. A particular focus in the project lies on communication between animated characters that exhibit credible personality traits and affective behaviour. The key challenge of the project is the fruitful combination of different research strands including situation-based generation of natural language and speech, semiotics of non-verbal expression in situated social communication, and the modelling of emotions and personality." The i-Guide showcase was heavily based on the NECA architecture.

Website: http://www.oefai.at/NECA/project/project.html

Text-to-Speech systems (TTS)

Name	DESCRIPTION
MBROLA	A free Phoneme-to-Speech system which includes many voices, among them English, French and German, for which a free Text-to-Phoneme is also available http://tcts.fpms.ac.be/synthesis/mbrola/
FreeTTS	A free TTS. "FreeTTS is a speech synthesis system written entirely in the JavaTM programming language. It is based upon Flite: a small run-time speech synthesis engine developed at Carnegie Mellon University. Flite is derived from the Festival Speech Synthesis System from the University of Edinburgh and the FestVox project from Carnegie Mellon University". http://freetts.sourceforge.net/
Festival	A free TTS. "Festival offers a general framework for building speech synthesis systems as well as including examples of various modules. As a whole it offers full text to speech through a number APIs: from shell level, though a Scheme command interpreter, as a C++ library, from Java, and an Emacs interface. Festival is multi-lingual (currently English (British and American), and Spanish) though English is the most advanced". http://www.cstr.ed.ac.uk/projects/festival/
Flite	A free TTS. "Flite (festival-lite) is a small, fast run-time synthesis engine developed at CMU and primarily designed for small embedded machines and/or large servers. Flite is designed as an alternative synthesis engine to Festival for voices built using the FestVox suite of voice building tools". http://www.speech.cs.cmu.edu/flite/
ViaVoice	A multilingual commercial TTS. http://www.scansoft.com/viavoice/
Nuance	A multilingual commercial TTS. http://www.nuance.com/

Markups [Pirker and Krenn, 2002]

Name	DESCRIPTION
VHML	"A markup language for the representation of different aspects of avatars, such as speech production, facial and body animation, emotional representation, dialogue management, and hyper and multimedia information". http://www.vhml.org
MPML	"MPML (Multimodal Presentation Markup Language) is an XML-based markup language developed to enable the description of multimodal presentation on the WWW based on animated characters". http://www.miv.t.u-tokyo.ac.jp/MPML/en/
SSML	"The essential role of the markup language is to provide authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. across different synthesis-capable platforms". http://www.w3.org/TR/speech-synthesis/
SAPI TTS	The Microsoft TTS markup. http://www.microsoft.com
SML	The sub-part of VHML concerned with the markup for speech synthesis.
RRL	"A Rich Representation Language for the description of agent behaviour in the NECA project". http://www.oefai.at/NECA/RRL/RRL_docs/RRL-FINAL.pdf

REFERENCES

[Allen, 1995] Allen, J. (1995). *Natural Language Understanding*. Benjamin Cummings.

[Armstrong-Warwick, 1993] Armstrong-Warwick, S. (1993). Preface to the Special Issue on Using Large Corpora. *Computational Linguistics*, 19(1):iii-iv.

[Bahl et al., 1983] Bahl, L., Jelinek, F., and Mercer, R. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Int., pages 179-190.

[Biber et al., 1999] Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). Longman Grammar of Spoken and Written ENGLISH. Pearson Education Limited, Edinburg, UK.

[Charniak et al., 1993] Charniak, E., Hendrickson, C., Jacobson, C., and Perkowitz, M. (1993). Equations for Part-of-Speech Tagging. Proceedings of the *Eleventh National Conference on Artificial Intelligence*, pages 784-789.

[Chu-Carroll and Brown, 1998] Chu-Carroll, J. and Brown, M. K. (1998). An Evidential Model for Tracking Initiative in Collaborative Dialogue Interactions. *User Modeling and User-Adapted Interaction*, 8:215-253.

[Clarkson and Rosenfeld, 1997] Clarkson, P. and Rosenfeld, R. (1997). Statistical Language Modeling using the CMU-Cambridge Toolkit. *Proc. Eurospeech 1997*, pages 2707-2710.

[Collins, 1997] Collins, M. (1997). Three Generative, Lexicalized Models for Statistical Parsing. In Cohen, P. R. and Wahlster, W., editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 16-23, Somerset, New Jersey. Association for Computational Linguistics.

[Core et al., 2003] Core, M. G., Moore, J. D., and Zinn, C. (2003). The Role of Initiative in Tutorial Dialogue. *10th Conference of the European Chapter of the Association for Computational Linguistics*.

[Klein et al., 2001] Klein, A., Schwank, I., Généreux, M., and Trost, H. (2001). Evaluating Multi-modal Input Modes in a Wizard-of-Oz Study for the Domain of Web Search. *Ann Blandford, Jean Vanderdonckt and Phil Gray (eds), People and Conputer XV - Interaction without Frontiers: Joint Proceedings of HCI 2001 and IHM 2001*, pages 475-483.

[Litman, 1996] Litman, D. (1996). Cue Phrase Classification using Machine Learning. *Journal of Artificial Intelligence Research*, 5:53-95.

[Manning and Carpenter, 2000] Manning, C. and Carpenter, B. (2000). Probabilistic Parsing Using Left Corner Language Models. In Bunt, H. and Nijholt, A., editors, *Advances in Probabilistic and Other Parsing Technologies*, pages 105-124. Kluwer Academic Publishers.

[M-PIRO] http://www.ltg.ed.ac.uk/mpiro/

[NECA] http://www.oefai.at/NECA/

[Ng and Zelle, 1997] Ng, H. T. and Zelle, J. (1997). Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing. *Special Issue on Empirical Natural Language Processing{AI Magazine Winter 1997*, 18(4):45{64.

[Pirker and Krenn, 2002] Hannes Pirker and Brigitte Krenn (2002), *Assessment of Markup Languages for Avatars, multimedia and multimodal system.* Neca Deliverable D9c, ÖFAI, http://www.oefai.at/NECA/publications/publication_docs/d9c.pdf

[Rabiner, 1989] Rabiner, L. (1989). A Tutorial on hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257-286.

[Whittaker and Stenton, 1988] Whittaker, S. and Stenton, P. (1988). Cues and Control in Expert-Client Dialogues. *Proc. of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 123-130.
[WYSIWYM] http://www.itri.bton.ac.uk/projects/WYSIWYM/wysiwym.html