

# Using Corpus-based Methods for Spoken Access to News Texts on the Web

*Alexandra Klein*

Austrian Research Institute  
for Artificial Intelligence  
Vienna, Austria  
alexandra -at- oefai.at

*Harald Trost*

Department of Medical Cybernetics  
and Artificial Intelligence  
University of Vienna, Austria  
harald -at- ai.univie.ac.at

## Abstract

The system described in this paper relies both on a multimodal corpus and a written newspaper corpus for processing spoken and written user requests to Austrian news texts. Requests may be spontaneous spoken and written utterances as well as mouse clicks; user actions may concern actual search, but also control of the browser. Because of spontaneous utterances, a large vocabulary and multimodal interaction, interpreting the user request and generating an appropriate system response is often difficult. Apart from a controller module, the system uses data from two corpora for compensating the difficulties associated with the scenario. Multimodal user actions, which were collected in Wizard-of-Oz experiments, serve as a base for the identification of patterns in users' spontaneous utterances. Furthermore, news documents are used for obtaining background knowledge which can contribute to query expansion whenever the interpretation of users' utterances encounters ambiguity or underspecification concerning the search terms.

## 1. Introduction

In this paper, we are going to describe the corpus base for a system which allows users to access news texts on the web by means of spontaneous speech, typed utterances, mouse clicks, or any combination thereof. The system described in this paper combines graphical with written and spoken input to access texts in Austrian newspapers that are available online, and it has to allow for queries addressing browsing functionality, structure and content by providing analysis mechanisms for distinguishing between these different levels. In order to interpret user actions in such a multimodal setting, the system has to rely on empirical data. For our system, we have compiled corpora containing both user actions and web documents for the search. An evaluation of the user-action corpus, which was obtained in Wizard-of-Oz experiments, led us to the development of a layered architecture for analyzing users' intentions. The analysis of the search terms in user requests is supported by contextual knowledge and world knowledge, which are derived from the document corpus with rule-based methods.

## 2. System and architecture

Several systems have incorporated actual speech browsing: In SLAM [1], it is possible to access prepared hypertext documents: pages, titles and hyperlinks can be activated by voice. Sam [2] provides the opportunity to utter browser commands, link names, and to a certain degree, names of entities defined on 'smart pages'. WebGALAXY [3] is a web extension of MIT's spoken dialogue system GALAXY. The project SALSA [4] also

aims at a purely speech-based web browser. The emphasis at the moment is on speech recognition problems, namely the recognition of unknown words and the effects of a dynamic vocabulary. SLAM, Sam and SALSA are realized for English whereas WebGALAXY is designed as a multilingual system. In this context, several commercial products also offer speech access to browsers.

In contrast to most of the cited system, we did not concentrate on speech recognition. Instead, we used a commercial, speaker-dependent system. As was mentioned earlier, another difference lies in the fact that our system allows users to formulate complex queries referring to content, document structure and the display in the browser. Therefore, our system architecture includes a controller, as can be seen in fig. 1 which gives an overview of the architecture.

The controller

- handles requests by voice, by written input or by mouse click,
- is responsible for task integration [5],
- coordinates consultation of knowledge bases,
- is in charge of output selection, and
- arranges storage of activated information [6].

The controller is also responsible for the interface between the language understanding component, which analyzes the users' utterances and extracts the action requests associated with it, and the browser-related module, which determines and carries out the requested action and updates the interaction protocol.

### 2.1. NLP components

A shallow processing module proceeds in two steps: patterns for search (e.g. *I am looking for an article about...*, *I want to start a new search...*), dates (e.g. *...in yesterday's newspaper...*) and newspaper names are identified and mapped (with the help of the controller) onto types of user action or commands for accessing the newspaper databases. After this preprocessing stage, the remaining words are considered search terms, and they are grouped according to grammatical relations and co-occurrence information.

### 2.2. News texts

As the system has been designed for online news texts from Austrian newspapers, it has to handle the difficulties associated with the domain of news, including the relatively wide range of topics covered (resulting in a large vocabulary and the frequent

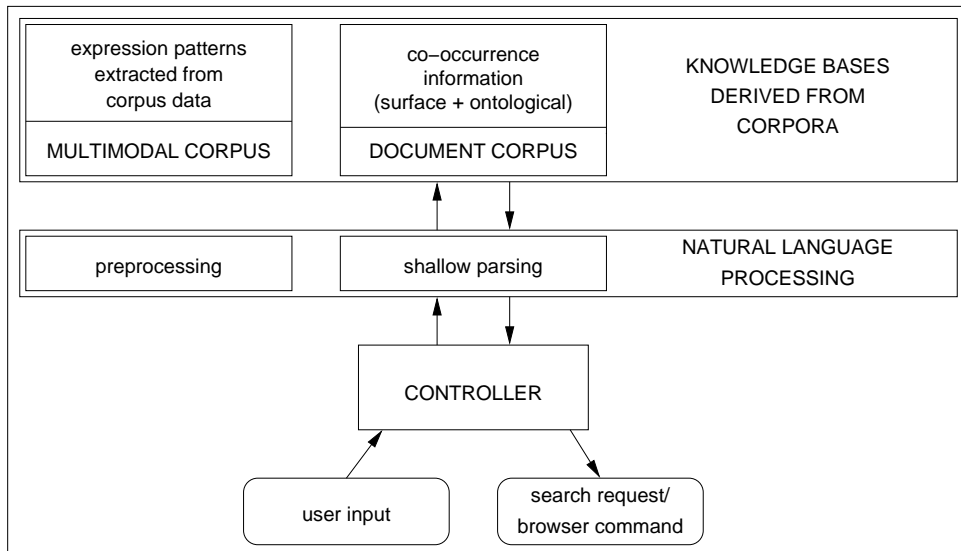


Figure 1: Architecture and knowledge sources.

occurrence of unknown words). A corpus of Austrian newspaper texts which has been collected during the past year gives an idea of the size and fluctuation of the vocabulary.

### 3. Multimodal corpus

#### 3.1. Experiments

The multimodal corpus was obtained with the help of Wizard-of-Oz (WoZ) experiments [7]. 43 test persons were given twelve tasks each, with the tasks being assigned different types and combinations of input modes: spoken input, typed input and free typed or spoken input. Both in the spoken and written input modes, the test persons were encouraged to mix browser commands with the actual search requests. According to their experience with the web, the users were grouped as experts and non-experts. During the experiments, it turned out that the non-experts in particular were comfortable combining the browser commands with actual search requests, i.e. *back to Politics*, and the use of speech influenced performance times and user satisfaction favorably. For the experiments, predefined web pages were used in order to achieve comparable results.

As far as speech recognition is concerned, we are well aware of the problems associated with WoZ experiments in a speech environment, particularly the simulation of speech recognition errors. Due to the large vocabulary, a speaker-dependent system was needed. Since there was no time to train 43 test persons on the speaker-dependent systems, speech recognition had to be simulated by the human wizard.

The four performance dimensions were task completion (determined in the evaluation), task completion speed (in seconds), complexity of interaction (in number of interaction per task) and the user response to the system (assessed by means of questionnaire-based satisfaction ratings). Results show that experts and non-experts prefer multimodal interaction over single input modes, and non-experts in particular were able speed up task completion times significantly with a combination of spoken and written input with mouse clicks [8].

Table 1 shows general results of the experiments. In summary, experts and non-experts were faster using all three input

modes.

Table 1: From the results of the WoZ experiments: number of interactions and time used.

|                         | spoken + mouse | written + mouse | spoken + written mouse |
|-------------------------|----------------|-----------------|------------------------|
| av. nr. of interactions |                |                 |                        |
| non-experts             | 24             | 22              | 22                     |
| experts                 | 26             | 26              | 25                     |
| av. nr. of seconds      |                |                 |                        |
| non-experts             | 553            | 1000            | 514                    |
| experts                 | 593            | 758             | 427                    |

#### 3.2. Corpus

The subjects' utterances were recorded, and typed input and mouse clicks were logged and later aligned with the spoken input.

Using the information obtained from the experiments has turned out to be a good way to predict the type and content of user actions. The speech corpus has helped in improving the quality as well as coverage of the speech understanding components. The improvements concern two areas in particular: the derivation of patterns for the preprocessing stage and the handling of characteristics of spontaneous speech. As far as patterns are concerned, the corpus of utterances spoken and typed by the 43 test persons resulted in a good overview of how different user actions are lexicalized, how people prefer to refer to dates and newspapers, and how they combine the different elements in their utterances. In the informal testing, which was carried out parallel to the development of the system, hardly any new forms had to be incorporated. Furthermore, the data from the experiments helped gain insight into spontaneous-speech phenomena which are likely to be encountered in the scenario.

Utterances are often elliptic and ungrammatical according to the standards of formal written language. False starts and repairs are also frequent. Morphological endings tend to be blurred or omitted, which often leads to errors in speech recognition. In many cases, particles can be used as cues for embedded metalinguistic information in utterances.

These findings were important as user habits and preferences must adequately be taken into account in system design to add value to an interface by adding spoken language [9]. Since there are no corpora for spoken and typed web access to news texts in German, the corpus has been a valuable resource for the development of the system.

### 3.3. Contrastive functionality

Apart from the adaptation of the system to user needs and the creation of a corpus, usability testing in a multimodal setting aims at gaining insight into the role of the speech input [10, 11]. In many scenarios, little is known about mode preference and coordination. Generally, it is agreed that a contrastive functionality [12] of several input modes including speech results in the highest performance as well as user acceptance since the naturalness and effectiveness of human language is based on its flexibility to support every possible way of interaction. Direct manipulation is still regarded as the most efficient, natural and intuitive way of interacting with objects on the screen, and mouse clicks are thus often used for simple tasks like following a link or moving between two documents. Typed input is often used for search terms as this medium is less error-prone than spoken language (in spite of typological errors). Spoken language tends to be employed for complex queries as they can be expressed quickly in this medium.

### 3.4. Typology of user actions

In the evaluation of the multimodal corpus, we have decided to distinguish between three types of user action:

- Combined queries referring to form and content, i.e. utterances which contain browsing requests as well as search terms, e.g. *back to the article about Tiger Woods in the Sports section*.
- Content queries containing only search terms, e.g. *information about the volleyball world championship*.
- Browsing requests and metalanguage, e.g. *back, I want to start a new search*.

## 4. Document corpus

For the analysis of the search terms in the users' queries, we have compiled a written corpus which contains a year's worth of articles from the Austrian daily newspaper *Der Standard*<sup>1</sup>. This corpus has served as a base for 'query expansion' in the sense that users' queries can be complemented with information lost during speech recognition, omitted by the users or linked to some contextual information neglected in the analysis. Apart from a robust natural-language understanding component, a sound representation of context and world knowledge considerably facilitates the interpretation of user requests. In our system, the document corpus serves as a foundation for deriving ontological knowledge which resolves contextual ambiguities and contradictions in the user utterances.

<sup>1</sup><http://derstandard.at>

### 4.1. Adjective-noun combinations

While the multimodal corpus has provided a list of keywords which refer to commands like 'new search', to browser commands, to newspaper names or sections of newspapers, or to dates, information relevant for the treatment of the search terms comes from the data in the newspaper corpus. From the corpus, pairs and counts of adjective/noun and adjective/proper occurrences were extracted [13]. These pairs were then stored to be used in the processing of the search terms. Since the newspaper corpus is automatically tagged, a considerable number of errors can be found in the list of pairs. We have therefore introduced a threshold: only adjective/noun or adjective/proper name pairs which have appeared a certain number of times at minimum are considered. This is feasible for large amounts of corpus data and reduces the number of errors in the interpretation of the search request, although naturally, systematic tagging errors cannot be detected by just taking into account frequency.

In order to neutralize inflectional forms, a very robust stemming algorithm reduces the adjective/noun pairs to an approximate 'stem'. Whenever an adjective/noun pair or an adjective/proper name pair is encountered among the search terms which are left after preprocessing has eliminated the features of metalanguage, the pair is processed by the stemming algorithm. If the 'base' form of the pair can be found in the repository of adjective/noun and adjective proper name combinations, it is passed on to the search engine as a query. The same is done with combinations of adjectives and nouns or proper names which can be reconstructed from previous contexts.

This approach separates typical phrases from specific lexicalizations. For example, the combination *Austrian ski star* will be considered while *last year's ski star* is only kept if it occurs frequently in the corpus. This way, the time reference can be eliminated for a search in older articles.

### 4.2. Ontological knowledge

Apart from the lexical co-occurrence relations described in the previous section, we have increasingly started to use the co-occurrence of entities, particularly named entities, in newspaper articles. As it is not feasible to use GermaNet [14] on large vocabularies as they occur in the news domain, deriving structured entities-and-relationship sets from corpora can help assign meaning to the texts in the corpora (and consequently, also to the query terms). This approach is similar to question-answering systems, e.g. [15], but in our case, the extraction of meaning from text is not carried out for finding the optimal answer, but rather for resolving ambiguities and underspecifications which occur in users' multimodal typed and spoken utterances. This kind of background knowledge helps the controller to place the user action in the appropriate context.

Knowledge about entities and their relations is extracted by means of rules. The resulting predicates are considered our basic 'ontological' database. Examples for the relations for persons can be found in table 2:

The corpus contains indications which can be used for determining the appropriate relations, e.g. if *former* is encountered together with a job title, this is an indicator for a *was-the* or *was-a* relation. These relations can be used as a 'Thesaurus' for query expansion which means that when a search term as it was uttered by the user (and understood by speech recognition if it was a spoken utterance) would yield too many results in the search process, it can be specified further by means of the relations. The same approach is used for locations, i.e. the relation between locations, the relation between

Table 2: *Example relations.*

| relation | example  |
|----------|--|
| is-the   | is-the(Romano Prodi,<br>European commission president) |
| was-the  | was-the(Margaret Thatcher,<br>British Prime Minister)  |
| is-a     | is-a(Michael Schumacher,<br>Formula 1 driver)          |

locations and persons, events etc. In this case, the relations reflect more a general ontology than a Thesaurus.

## 5. Conclusion

It has been shown that corpora can support the interpretation of user requests in a multimodal system, e.g. when the user input appears to be underspecified or if it seems to contradict earlier assumptions of the system about the interaction. The multimodal corpus helps determine the users' intentions by comparing the utterances to previously encountered patterns. Since the meaning of utterances in our domain can be 'layered', i.e. it can refer to different aspects of interaction, namely search, browsing commands and the state of the system, the preprocessing stage can map typical patterns of usage to the appropriate types of action. The corpus of news documents provides background knowledge for the interpretation of the search terms. By means of lexical co-occurrence information as well as simple ontological relations which have both been extracted from the corpus, important search terms can be identified, grouped and combined with terms supplied by lexically or ontologically related information.

The design of multimodal systems often remains difficult for scenarios with complex types of interaction and domains with large vocabularies - which make interaction more natural for the user. Yet, we have come to the conclusion that the use of corpora not only for language models, where they are traditionally employed, but also for a meaning interpretation of user requests can lead to more robust and usable systems.

## 6. Acknowledgements

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture; the project is supported by the Austrian Science Fund (P13704-INF). The authors would like to thank Michel Génèreux and Ingrid Schwank for their contribution to the Wizard-of-Oz experiments.

## 7. References

- [1] D. House, "Spoken-Language Access to Multimedia (SLAM)," Master's thesis, Dept. of Computer Science and Engineering, Oregon Graduate Institute, Beaverton, OR, April 1995, also appears as Technical Report TR 95-008 and URL <ftp://speech.cse.ogi.edu/pub/docs/SLAM-thesis.ps.Z>.
- [2] C. T. Hemphill and P. R. Thrift, "Surfing the Web by Voice," in *Proc. ACM Multimedia*, San Francisco, CA, November 1995, pp. 215-222.
- [3] R. Lau, G. Flammia, C. Pao, and V. Zue, "WebGALAXY - Integrating Spoken Language and Hypertext Navigation," in *Proceedings of EUROSPEECH 1997*, Rhodes, Greece, September 1997, pp. 883-996.
- [4] P. Fung, C. s. Shun, L. K. Leung, L. W. Kat, and L. Y. Yee, "SALSA version 1.0, A speech-based web browser," in *ICSLP 98: Fifth International Conference on Spoken Language Processing*, vol. 4, Sydney, Dec. 1998, pp. 1615-1619.
- [5] M. A. Grasso and T. Finin, "Task Integration in Multimodal Speech Recognition Environments," *Crossroads*, vol. 3, no. 3, pp. 12-22, Spring 1997.
- [6] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, "Unification-based Multimodal Integration," in *Proceedings 35th Annual Meeting of the ACL*. Madrid: ACL Press, July 1997.
- [7] N. M. Fraser and G. N. Gilbert, "Simulating speech systems," *Computer Speech and Language*, vol. 5, no. 1, pp. 81-99, 1991.
- [8] A. Klein, I. Schwank, M. Génèreux, and H. Trost, "Evaluating Multimodal Input Modes in a Wizard-of-Oz Study for the Domain of Web Search," in *People and Computer XV - Interaction without Frontiers: Joint Proceedings of HCI 2001 and IHM 2001*, A. Blandford, J. Vanderdonck, and P. Gray, Eds. Springer: London, September 2001, pp. 475-483.
- [9] N. Yankelovich, "Using natural dialogs as the basis for speech interface design," in *Automated Spoken Dialogue Systems*. MIT Press, 1997 to appear.
- [10] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, and I. Smith, "The efficiency of multimodal interaction: A case study," in *Proceedings of the ICSLP*, Sydney, Australia, December 1998.
- [11] D. L. Litman, S. Pan, and M. A. Walker, "Evaluating response strategies in a web-based spoken dialogue agent," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. II. Montreal, Quebec, Canada: Université de Montreal, August 10-14 1998, pp. 780-786.
- [12] S. L. Oviatt, "Integration Themes in Multimodal Human-Computer Interaction," in *Proceedings of the ICSLP*, vol. 2. Yokohama: Acoustical Society of Japan, 1994, pp. 551-554.
- [13] A. Klein, Puig-Waldmüller, and H. Trost, "Robust interpretation of user requests for text retrieval in a multimodal environment," in *COLING 2002: Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, 2002, pp. 1233-1237.
- [14] C. Kunze and L. Lemnitzer, "GermaNet - representation, visualization, application," in *Proceedings LREC*, vol. V, 2002, pp. 1485-1491.
- [15] U. Hermjakob, "Parsing and question classification for question answering," in *Proceedings of the Workshop on Open-Domain Question Answering, Association for Computational Linguistics (ACL), Toulouse, France, 2001*.