Corpus-based Research

Friday 9 June 2006, Big conference room, Max Planck Institute for Psycholinguistics, Nijmegen

Organized by the *Nederlandse Vereniging voor Fonetische Wetenschappen*

PROGRAMME

9.30 - 9.55	**Welcome - Coffee & tea**
9.55	**Opening**
10.00	**Modified repeats: one method for asserting primary rights from second position** *Tanya Stivers*
10.30	**Audiovisual cues to a speaker's confidence level** *Marc Swerts*
11.00	**Coffee & tea**
11.30	**Automatic phonetic transcription of large speech corpora: a comparative study** *Christophe Van Bael*
12.00	**Morphological information and acoustic duration in Dutch compounds** *Victor Kuperman*
12.30	*Meeting of NVFW members*
13.00 - 14.00	**Lunch**
14.00	**Example-based large vocabulary recognition** *Dirk Van Compernolle*
14.30	**Multimedia retrieval** *Arjan van Hessen*
15.00	**Coffee & tea**
15.30	**Methodologies for improving the g2p conversion of Dutch names** *Henk van den Heuvel*
16.00	**Text-induced spelling correction** *Martin Reynaert*
16.30	**Drinks**

(see also: http://www.fon.hum.uva.nl/FonetischeVereniging/)

De lezingendag zal plaatsvinden in de grote conferentiezaal van het Max Planck Instituut voor Psycholinguïstiek te Nijmegen. Deze zaal vindt u links achter in de entreehal van het instituut.

Tijdens de lunchpauze zullen broodjes, soep, fruit, en verschillende soorten dranken verkrijgbaar zijn in de kantine van het Max Planck Instituut.

Het Max Planck Instituut is met het openbaar vervoer goed te bereiken vanaf het Centraal Station:

- Bus 4 naar Wychen
- Bus 6 naar de Brabantse Poort (*niet* naar Neerbosch!)
- Bus 10 Heyendaal shuttle
- Bus 54 naar Grave
- Bus 83 naar Venlo

(zie ook: http://www.mpi.nl/world/maps/uni.html)

Het Centraal station is een gewone halte voor deze buslijnen, en geen eindpunt. Zorg er daarom voor dat u niet alleen in een bus stapt met het goede nummer, maar ook met de goede richting. Stap uit bij Tandheelkunde. Loop in de rijrichting van de bus. U komt dan bij een T-kruising. Volg de Erasmuslaan naar links. Neem dan de eerste zijstraat rechts. U loopt langs de aula van de Radboud Universiteit en langs een slagboom. Aan het eind van de parkeerplaats ziet u door de bomen een wit stenen gebouw. Dit is het Max Planck Instituut. Volg het voetpad langs het gebouw. Het brengt u naar de ingang. Zie ook de kaart van de RU-campus met daarop het Max Planck Instituut op http://www.mpi.nl/world/maps/uni.html.

Het Max Planck Instituut bezit een eigen parkeerplaats waar gasten gratis kunnen parkeren.

Voor nadere informatie kunt u terecht bij:

Mirjam Ernestus

Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen
The Netherlands
tel: +31-24-3612970

Abstracts

Modified repeats: one method for asserting primary rights from second position

Tanya Stivers
Max Planck Institute for Psycholinguistics, Language & Cognition Group

This presentation examines one practice speakers have for confirming when confirmation was not otherwise relevant. The data are a collection taken from video and audio recordings of spontaneous face-to-face and telephone conversations between family members and friends. The practice I analyze here involves a speaker repeating an assertion previously made by another speaker in modified form with stress on the copula/auxiliary. It is argued that these modified repeats work to undermine the first speaker's default ownership and rights over the claim being made and instead assert the primacy of the second speaker's rights to make the statement.
Two types of modified repeats are identified: partial and full. Though both involve competing for primacy of the claim, they occur in distinct sequential environments: the former are generally positioned after a first claim was epistemically downgraded whereas the latter are positioned following initial claims that were offered straightforwardly, without downgrading.

Audiovisual cues to a speaker's confidence level

Marc Swerts
Communication and Cognition, Tilburg University

I will present the results of a number of experiments on the role of audiovisual prosody for signalling and detecting meta-cognitive information in question-answering. The first study consists of an experiment in which participants are asked factual questions in a conversational setting, while they are being filmed. Statistical analyses bring to light that a speakers' confidence level is cued by a number of visual and verbal properties. Interestingly, it appears that answers tend to have a higher number of marked auditive and visual feature settings, including divergences of the neutral facial expression, when a speaker's confidence level is low, while the reverse is true for non-answers. The second study is a perception experiment, in which a selection of the utterances from the first study is presented to participants in one of three conditions: vision only, sound only or vision+sound. Results reveal that human observers can reliably distinguish high confidence responses from low confidence responses in all three conditions, be it that answers are easier than non-answers, and that a bimodal presentation of the stimuli is easier than the unimodal counterparts. The talk will end with some perspectives on related work about difference in the expression of confidence level between speakers who differ in age and cultural background.

Automatic phonetic transcription of large speech corpora: a comparative study

Christophe Van Bael
CLST, Radboud University Nijmegen

In a recent study, we investigated whether automatic transcription procedures can approximate manually verified phonetic transcriptions typically delivered with contemporary large speech corpora. Ten automatic procedures were used to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The resulting transcriptions were compared to manually verified phonetic transcriptions from the same corpus.

We found that signal-based procedures could not approximate the manually verified phonetic transcriptions. A knowledge-based procedure did not give optimal results either. Quite surprisingly, a procedure in which a canonical transcription, through the use of decision trees and a small sample of manually verified phonetic transcriptions, was modelled towards the target transcription, performed best. The number and the nature of the remaining discrepancies compared to inter-labeller disagreements reported in the literature. This implies that future corpus designers should consider the use of automatic transcription procedures as a valid and cheap alternative to expensive human experts.

Morphological information and acoustic duration in Dutch compounds

Victor Kuperman
Radboud University Nijmegen

Recent literature demonstrates that articulatory salience in speech (e.g. acoustic duration and loudness) is sensitive to the amount of information carried by phonemes, syllables and words. The more predictable (i.e. less informative) a linguistic unit is in its lexical or phonological environment, the less salient its realization. Examples of this phenomenon, especially common in spontaneous speech, include acoustic reduction of highly frequent functional words, durational shortening or deletion of predictable discourse markers, and longer articulation of phonemes with higher contribution to word recognition.
We tested whether the amount of information supplied by morphological units adds to other (phonetic, prosodic and lexical) domains of predictability and modulates the acoustic duration of affixes. This research focused on the interfixes -s- or -e(n)- in Dutch compounds. The selection of the interfix is not determined by rules, but depends on probabilistic characteristics of the left and right constituent families (sets of compounds sharing the left/right constituent with the target). The goal was then to detect the impact of families in the interfix articulation. The study was based on two datasets collected from the "Library of the Blind" component of the Spoken Dutch Corpus: 1156 tokens containing the interfix -s- and 787 tokens containing the interfix -e(n)-. The dependent variables of the study were acoustic durations of the interfixes, and, for the interfix -e(n)-, the number of segments in the interfix. The acoustic duration of phonemes was determined with the help of an ASR, while the presence of [n] in the interfix was established by two phoneticians. We report the correlation of acoustic salience of the interfix and the amount of information in both positional families, as well as the distribution of interfixes in the left family. Moreover, we demonstrate that a number of durational effects induced by phonetic and prosodic factors and so far only observed under laboratory conditions is also found in the genre of lively read aloud speech.

Example-based large vocabulary recognition

Dirk Van Compernolle
ESAT/PSI, K.U. Leuven

Hidden Markov Models(HMM) have dominated speech recognition for over two decades. HMMs are an embeddiment of a beads on a string model in which a sentence is a sequence of words, a word a sequence of phonemes and a phoneme a sequence of states. An HMM-state (in the acoustic model) models a sub-phonetic speech fragment as a short-time stationary event. HMMs have great advantages: the concept is straightforward and the parameters in the model are trained from data available in large databases. Moreover HMMs have proven to be extremely scaleable: larger database allow for more detailed models with more parameters while more powerful CPUs make it possible to use these more detailed models in real-time systems. The success of HMMs has been the single most important driving force in the use of large databases and statistical techniques in the field of speech and language.
Nevertheless HMMs are far from ideal in their speech modeling concept. Especially the short-

time stationarity assumption is contradictory to the nature of speech which often looks more like a concatenation of transients than a concatenation of stationary segments. In order to overcome these fundamental weaknesses a new line of speeech recognition systems is currently being developed that avoids the modeling step all together and does recognition straight from the data by the application of template matching. This avoids the step of imperfect modeling and at the same time itis in line with recent psycholinguistic findings that claim that many individual traces of speech fragments are permanently stored in memory. Template based systems require that the full database is accessible at recognition time; which thanks to further increases in hardware performance is almost within reach. However, template based recognition has fundamental weaknesses as well: it relies on the score of one or a few examples only to compute a distance score.
In this presentation we will compare the pro's and con's of HMM and template based recognition. Both of the them could not exist without the availability of large corpora of speech. However, the way in which these corpora are used in an actual recognition system are drastically different for both methods.

Multimedia retrieval

Arjan van Hessen
HMI, University of Twente

The number of digital multimedia collections is growing rapidly. Due to the ever declining costs of recording audio and video, and due to improved preservation technology, huge data sets containing text, audio, video and images are created, both by professionals and non-professionals.
The reasons for building up these collections may vary. Organisations such as broadcast companies consider the production and publishing of multimedia data as their core business. Within these companies there is a tendency to search for "means" to get more out the produced content: a nice example is the added basic search functionality in the "uitzending gemist" collection. Other organisations are merely interested in obtaining insight in the internal information flow, for internal (corporate meetings that are recorded) or public use (council meetings that are recorded and webcasted). A number of organisations in the Netherlands administer spoken-word archives: recordings of spoken interviews and testimonies on diverging topics such as retrospective narratives, eye witness reports and historical site descriptions. Modern variants of these spoken-word archives are archives of 'Podcasts', 'Vodcasts' (video podcasts) and 'Vlogs' (video weblog), created in order to share 'home-made' information with "the world".
The Human Media Interaction (HMI) group is set within the computer science department and the Centre of Telematics and Information Technology (CTIT) and has a long history in multimedia retrieval research. Especially the use of audio mining and speech recognition technology in multimedia retrieval (SDR or spoken document retrieval) is an important research focus.
The presentation is focussed on the possibility to index and access spoken archives via the use of automatic speech recognition technology. The index, based on the imperfect recognition results is then used to search the document collection and relate individual documents to other information sources in (potentially) any media format. We will discuss the running demo application in which the recognised speech of the 8 o'clock news is used to connect news items with 5 (most) similar newspaper documents from the Twente News Corpus.

Methodologies for improving the g2p conversion of Dutch names

Henk van den Heuvel
CLST. Radboud University Nijmegen

Names pose particular problems for grapheme-to-phoneme (g2p) converters. This is due to their non-standard orthography caused by foreign origin or fossilisation of older spelling forms. In the Autonomata project a variety of techniques is studied to improve the g2p conversion of Dutch names, more specifically: first names, second names, street names and town names. In Autonomata, a standard g2p converter is augmented with a name-specific phoneme-to-phoneme (p2p) converter that captures the peculiarities of names. Based on large collections of names with a manually verified phonetic transcription, the p2p is trained with the specific information it requires. Various inductive and deductive approaches are studied to achive this goal. We will exemplify our approach by showing results on the g2p of Dutch first names.

Autonomata is carried out in the framework of the STEVIN-programme.
Partners in the project are the Radboud University Nijmegen, Ghent University, Utrecht University, Nuance, and TeleAtlas.

Text-Induced spelling correction

Martin Reynaert
Communication and Cognition, Tilburg University

In this talk we present an overview of our PhD-work on Text-Induced Spelling Correction. The work presents a novel approximate string matching algorithm for indexed text search. The algorithm is based on a hashing function which uniquely identifies strings composed of the same subsets of characters, i.e. anagrams, by means of a numeric value. The numeric value allows for searching for character strings differing from a particular string by a predefined number of characters. This forms an ideal basis for a novel spelling error detection and correction algorithm, which we call Text-Induced Spelling Correction or TISC. Our system uses nothing but lexical and word cooccurrence information derived from a corpus, a very large collection of texts in a particular language, to perform context-sensitive spelling error correction of non-words. Non-words are word strings produced unintentionally by a typist that deviate from a convention about how words are to be spelled in order to be considered real-words within the language. We will highlight the differences between our character-based similarity key and the language specific similarity keys as employed in, for instance, the well-known Soundex and Phonix phonetic spelling systems. The spelling error detection and correction mechanism we propose uses not only isolated word information, but also context information. It performs context-sensitive error correction by deriving useful knowledge from the text to be spelling checked. This enables our system to correct typos for which it does not have the correct word in its dictionary. Apart from this, some typos are ambiguous in that they may resolve into two or more different words. We investigate in depth the relationship between a typo and its context and propose a new algorithm for ranking correction candidates that specifically makes use of the typo's context.
We further discuss the tension between the wish of developers of spelling correction systems of catering for phonetic spelling errors and the cost of this in terms of the system's precision. Extensive evaluations on both English and Dutch allow us to illustrate this by discussing the performance of Aspell and the Microsoft Proofing Tools in this regard.