# Strings Of Natural Languages



Learning a second language is often difficult. One major reason for this is the way we learn: We try to translate the words and concepts of the other language into those of our own language. As long as the languages are fairly similar, this works quite well. However, when the languages differ to a great degree, problems are bound to appear. For example, to someone whose first language is French, English is not difficult to learn. In fact, he can pick up any English book and at the very least recognize words and sentences. But if he is tasked with reading a Japanese text, he will be completely lost: No familiar letters, no whitespace, and only occasionally a glyph that looks similar to a punctuation mark appears. Nevertheless, anyone can learn any language. Correct pronunciation and understanding alien utterances may be hard for the individual, but as soon as the words are transcribed to some kind of script, they can be studied and - given some time - understood. The script thus offers itself as a reliable medium of communication. Sometimes the script can be very complex, though. For instance, the Japanese language is not much more difficult than German - but the Japanese script is. If someone untrained in the language is given a Japanese book and told to create a list of its vocabulary, he will likely have to succumb to the task. Or does he not? Are there maybe ways to analyze the text, regardless of his unfamiliarity with this type of script and language? Should there not be characteristics shared by all languages which can be exploited? This thesis assumes the point of view of such a person, and shows how to segment a corpus in an unfamiliar language while employing as little previous knowledge as possible. To this end, a methodology for the analysis of unknown languages is developed. The single requirement made is that a large corpus in electronic form which underwent only a minimum of preprocessing is available. Analysis is limited strictly to the expression level; semantics are purposefully left out of consideration. This distinguishes this work clearly from other works, limits comparability to some extent, and may make detection of some kinds of language features hard or even impossible. Only unsupervised analysis is admissible, and no specific information on grammatical rules, ways to segment the text, what separators look like etc. is employed. Furthermore, no parameters such as absolute thresholds or the selection of a preset number of n-best candidates are allowed; all parameters and evaluation must be relative and

justifiable, not based on experimental results. Though this makes this thesis task harder, it also offers the advantage that parameters are not required, and thus need not be adjusted or optimized to fit to a corpus or language. This work examines English, German, Hebrew and Japanese. It presents ways to automatically create excerpts from a corpus, detect syntactic separators and segment text. EAN/ISBN : 9783836606271 Publisher(s): Diplomica Discussed keywords: Computerlinguistik, Korpuslinguistik Format: ePub/PDF Author(s): Stengel, Markus