Diplomarbeit

Markus Stengel

# Strings of Natural Languages

*Unsupervised Analysis and Segmentation on the Expression Level*

Diplom.de

**Bibliographic information published by the German National Library:**

The German National Library lists this publication in the National Bibliography; detailed bibliographic data are available on the Internet at http://dnb.dnb.de .


ISBN: 9783836606271

**Markus Stengel**

# Strings of Natural Languages

## Unsupervised Analysis and Segmentation on the Expression Level

Diplom.de

Diplomarbeit

ISBN: 978-3-8366-0627-1

Markus Stengel

# Strings of Natural Languages

*Unsupervised Analysis and Segmentation on the Expression Level*

Diplom.de

Markus Stengel
**Strings of Natural Languages**
Unsupervised Analysis and Segmentation on the Expression Level

ISBN: 978-3-8366-0627-1
Druck Diplomica® Verlag GmbH, Hamburg, 2008
Zugl. Eberhard-Karls-Universität Tübingen, Tübingen, Deutschland, Diplomarbeit, 2006

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# List of Abbreviations

| | |
|---|---|
| **AIC** | algorithmic information content |
| **IC** | index of coincidence |
| **ID** | identification (number) |
| **KCC** | Kolmogorov-Chaitin complexity |
| **MR** | meta-rating |
| **PSR** | prefix, suffix, and/or reduced segmentation unit |
| **RLE** | run-length encoding |
| **SIC** | Shannon information content |
| **SOV** | subject-object-verb (sentence structure) |
| **SU** | segmentation unit |
| **SVO** | subject-verb-object (sentence structure) |
| **TF** | term frequency |
| **LCS** | longest common subsequence |
| **LZW** | Lempel-Ziv-Welch (compression) |
| **LZMW** | Lempel-Ziv-Miller-Wegman (compression) |
| **LZMW78** | my modification of LZMW (compression) |

# Introduction

*The limits of my language are the limits of my mind. All I know is what I have words for.* – Ludwig Wittgenstein

Everyone who has ever learned a second language knows how hard it is. There are always differences: Some are glaringly obvious, and others are so subtle that even their concepts are difficult to understand. 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 which we are comfortable with.

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 most 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?