

An On-Line Japanese Handwriting Recognition System integrated
into an E-Learning Environment for Kanji

Diplomarbeit

zur Erlangung des Grades
eines Diplom-Linguisten
der
Fachrichtung 4.7 Allgemeine Linguistik
der Universität des Saarlandes.

Anfertigt von Steven B. Poggel
sbp@coli.uni-saarland.de

unter Leitung von
Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster
und
Dr. Tilman Becker

Saarbrücken, den 31.03.2010

Contents

1	Introduction	3
1.1	Motivation	3
1.1.1	Integrating NLP and E-Learning	3

Chapter 1

Introduction

1.1 Motivation

In the history of Computational Linguistics there have been a several attempts to integrate natural language processing techniques with existing technologies. This work is one just like that. Concretely, we will try to create a handwriting recognition for Japanese Kanji. That seems interesting, because Kanji is an iconographic writing system, thus handwriting recognition (HWR) can follow different patterns than in alphabetical writing systems like latin.

Studying Japanese language is a complex task, because a new learner has to get used to a new vocabulary that - coming from a European language - has very little in common with the vocabulary of his mother tongue, unlike in European languages where quite often there are several intersections. The learner also needs to learn a new grammar system. Broadly speaking, most of the central European languages follow a subject-verb-object (SVO) structure. Japanese follows a subject-object-verb (SOV) structure therefore creating additional difficulty, comparable with German subclause structures that are a source of error for learners of German. Yet, the most notable difference for a language learner with a central European mother tongue is of course the writing system. The Japanese writing system uses three different scripts. The Kana scripts Hiragana and Katakana are syllabic, each character represents a syllable. Each syllable consists of either a vowel, a consonant and a vowel, or a consonant cluster and a vowel. Hiragana and Katakana represent roughly the same set of syllables and both have around 40-50 characters that can be modified with diacritics and thus yield additional syllable representations. Therefore, these scripts are a hurdle, but relatively unproblematic, due to their limitation in number of characters. Besides, they look quite distinct, so there is the problem of confusing one character with another, but this is limited to a relatively short period of learning those characters.

Kanji, however, is an iconographic writing system that has around 2000 characters, which are built up of around 200 subunits called 'radicals'. So one part of the complexity lies in the number of characters. The other part of the complexity lies in the general concept of representing an idea or concept with a character instead of representing the phonemes of the spoken language with graphemes in connection with some language specific pronunciation rules. Another difficulty lies in connecting the characters with their pronunciations. Most characters have multiple pronunciations and for a language learner, studying Japanese vocabulary is a double or triple task compared to languages using a Latin or at least an alphabetic writing system. Therefore, the two tasks of learning the Kanji and studying the vocabulary together can epitomise a very high learning curve. A subordinated issue connected to that is that quite often subjectively 'simple' vocabulary comes with complex Kanji. Some e-learning applications have taken on that issue by creating a learning environment in which a learner can connect learning vocabulary with studying the Kanji.

1.1.1 Integrating NLP and E-Learning

In this project, we would like to approach the issue of studying Kanji in an e-learning application. The novelty about it is a handwriting recognition that gives the learner the ability to actually practise writing the Kanji, instead of the rather limited multiple choice recognition that most other applications use.

References

**Document created on Wednesday 31st
March, 2010 at 03:19**