A SET OF JAPANESE WORD COHORTS
RATED FOR RELATIVE FAMILIARITY
Takashi Otake and Anne Cutler
Dokkyo University and Max-Planck Institute for Psycholinguistics

ABSTRACT

A database is presented of relative familiarity ratings for 24 sets of Japanese words, each set comprising words overlapping in the initial portions. These ratings are useful for the generation of material sets for research in the recognition of spoken words.

1. INTRODUCTION

Spoken-language recognition proceeds in time - the beginnings of words arrive before the ends. Research on spoken-language recognition thus often makes use of words which begin similarly and diverge at a later point. For instance, Marslen-Wilson and Zwitserlood\(^1\) and Zwitserlood\(^2\) examined the associates activated by the presentation of fragments which could be the beginning of more than one word - in Zwitserlood's experiment, for example, the fragment kapit- which could begin the Dutch words kapitein ('captain') or kapitaal ('capital'). Taft and Hambly\(^3\) compared recognition of English words beginning in the same way but ending differently such as difficult and diffident; Slowiaczek, Nusbaum and Pisoni\(^4\) presented words with initial overlap to listeners under conditions of noise masking. Many other experiments have used such material.

In our research on the role of pitch-accent pattern in the recognition of spoken Japanese words\(^5\)\(^-\)\(^7\) we were interested in the question of whether listeners make use of suprasegmental cues to word identity at an early stage in lexical processing. Therefore we selected pairs of words which begin with the same CVCC sequence but which, in Tokyo Japanese, have different accential patterns. One member of each pair had accent pattern 1, while the other had accent pattern 0, 2 or 3. Examples are nagasa versus nagashi, which both begin naga-.

In a gating experiment (see \(^6\), for details) these words were presented in fragments of increasing length to listeners who guessed what the words might be. The guesses to smaller fragments were naturally hardly ever correct. For instance, given na- from nagasa, listeners produced 16 different word guesses, all incorrect: Narita, Natoo, Nara, Naruse, Nasa, nanzan, nama, nasu, napukin, nabe, naifu, namida, naito, naka, nagashi, nakai.

There are many types of perceptual experiments in which listeners are asked to guess about the identity of a speech signal which is in some way difficult to perceive; in gating the input is fragmentary, but other methods involve presentation of filtered or noise-masked or faint signals. In most such studies there is a strong familiarity effect: listeners guess words which are familiar to them rather than words which are unfamiliar. The above list suggests that the same was true in this study. However, in order to establish that this was so, it was necessary to compare the relative familiarity of the guessed words and the actually presented words. Unfortunately we found no existing database available for such a comparison.

It was therefore necessary to collect familiarity ratings for the words in question. Studies of subjective familiarity rating (Gernsbacher\(^8\), Kreuz\(^9\)) have shown very high inter-rater reliability and a better correlation with experimental results in language processing than is found for frequency counts based on written text. We therefore collected relative familiarity judgements from 45 naive raters for every stimulus-response pair which our experiment had produced, as well as for the paired items which we had used as stimuli. Across the 48 stimulus words which we used, there were in total 1033 separate guesses, and we also collected ratings for the 24 paired sets of stimuli. Thus nagasa occurred in 26 pairs, because it was rated for relative familiarity with its stimulus pair nagashi, with the 15 words other than nagashi listed above which were produced as guesses given the fragmentina-, and for 10 further guesses which were produced at other fragments.

Because the word guesses naturally began with the fragment which listeners had heard, the words which are compared overlap in the initial portions. These familiarity ratings could therefore, we believe, prove considerably useful to other researchers conducting spoken-word recognition studies in Japanese, and it is for this reason that we make the results of our ratings study available in the present report.

2. METHOD

Subjects. Subjects were 45 undergraduate students at Dokkyo University and Tsuda College, who completed the questionnaire in return for a small payment. They were all native speakers of Japanese from the Kanto area.
Materials and procedure. 1147 pairs of words were printed on 6 pages of B4 size paper. Instructions were printed on the left side of the first page. Words were written (according to the relevant convention for their orthographical representation) in kanji, hiragana or katakana, or on occasion in a combination of these. All loanwords were written in katakana. Most words of Chinese origin were written in Kanji, but some Kanji words, which in the judgement of the first author were unlikely to be well known in their conventional written form, were written either in hiragana, or with additional kana to indicate pronunciation. The total of 1147 included all stimulus words from the gating study plus all guessed words from the gating study, with some additional filler items which, unlike the items derived from the gating study, did not form extensive initial-overlap sets. 12 pairs were repeated for use as a reliability check.

Subjects were tested individually or in a group of as many as 6 in a room. They completed the questionnaire at their own speed. Their instructions were: "The purpose of this investigation is to investigate the familiarity of Japanese words to Japanese people. You will see 1147 pairs of words on the following pages. Look at each pair, and decide which word is more familiar to you. Once you decide, circle the word you chose. You may feel that it is hard to decide, but never fail to choose one from each pair of words. For instance, you may see yama - kompyuutaa ('mountain-computer'). Look at these words, and decide which word is more familiar to you. If you feel that yama is more familiar to you, circle that word. If you feel kompyuutaa is more familiar, choose that word."

3. RESULTS AND DISCUSSION

The completed questionnaires were scored and the number of respondents choosing each member of each pair collated.

The repeated items used for the reliability check produced a correlation of .97 between the first and second occurrence.

The guessed words proved more familiar than the stimulus words in 62.62% of cases. For guesses with the same initial accent pattern (HL- or LH-, where LH- collapses across all accent patterns except accent 1), rated familiarity was higher than the stimulus word in 61.51% of cases, while for guesses with different initial accent pattern, rated familiarity was higher than the stimulus word in 64.96% of cases.

The raw data from the sets of words derived from the gating study, in terms of numbers of subjects rating each item as relatively more familiar than the stimulus word it was paired with, are listed in the Appendix. Accent pattern in Tokyo Japanese of the guessed words, which was the focus of our study, is not recorded in the Appendix since this information is easily available from dictionary sources.

The guesses provide a view of the range of lexical options available to speakers of Tokyo Japanese given the initial ambiguous word fragments used in the study. Perhaps predictably, there is a tendency for the guesses to be reasonably familiar and for the words used as stimulus material also to be regarded as familiar. Thus many of the guesses are rated as more familiar than the stimulus words by around half of the subjects. However, the lists do include some which are rated as more familiar than the stimulus words by most subjects (for nagasa, for example, such words are nasu, napukin, naifu, namida), as well as others rated as less familiar by most subjects (for nagasa, Naruse, nanzan, nakai, nagaka, nagase).

4. ACKNOWLEDGMENTS

This research was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture (Grant number: 06610475) and a grant from Max Planck Society, both to the first author. We thank Daisuke Ito and Takaaki Nakazawa (Dokkyo University), Keiko Noguchi (Tsuda College), Kiyoko Yoneyama, Aki Arai, Yoko Kuroda and Maki Yoshida (Ohio State University) for assistance. A full report of this work will appear in Dokkyo Studies in Data Processing and Computer Science, 18.

5. REFERENCES

8. Gernsbacher, M.A. “Resolving twenty years of


### 6. APPENDIX

The numbers after each item refer to the number of subjects rating that item as more familiar than the stimulus word. The remainder of the 45 subjects rated the stimulus word as more familiar. Thus the example *nagasa*, 24 subjects rated *nagasa* as more familiar than *nagashi*, while 21 subjects rated *nagashi* as more familiar than *nagasa*.

1. **bakuho** bakuchi 19, baasan 35, baggu 37, baiku 40, baiorin 38, baka 36, bakku 31, baku 5, bakudan 30, bakuro 37, banana 43, basya 25, bataa 41, bataa 37, batto 14

2. **bukuchi** bakuchi 26, bacchi 28, baggu 37, haka 36, bakemono 25, bakkin 31, bakku 35, bakunetto 12, baku 4, bakuchiku 18, bakudan 26, basuketto 34, batsu 35, batta 29, batto 32

3. **hanabi** hanawa 9, hada 31, hage 25, hagiuki 14, haisha 31, hakka 5, hakubi 5, hamachi 17, hamaki 9, haru 30, washi 25, hatena 11, hato 26, hattari 19, hatto 7, hachi 22

4. **hanawa** hanabi 36, hadagi 30, hande 30, hana 41, hanami 39, hanamizu 35, hanao 13, hanatsui 35, hanata 30, hanaya 36, hanga 25, hankachi 41, hanko 36, hankyou 34, hanran 27, hanishi 26, hantai 34, hanten 12, happa 36, haramaki 20, hari 12, hashi 37, hasu 21, hato 31, hatten 32

5. **hokubo** hokuro 35, hotaru 32

6. **hokuro** hokubo 16, hoken 20, hokkai 6, hokke 12, hokki 22, hokusai 22, hokuchuu 14, hokudan 26, basuketto 34, hatsu 35, batta 29, batto 32

7. **kamotsu** kamotsu 12, kamoku 28, kamon 13, kamoniku 14, kamoku 28, kamon 13, kamoniku 14, kan 18, kana 18, kanbotsu 5, kankou 37, kamuri 13, kanpachi 12, kanso 36, kantan 23, kanten 29, karasu 24, karate 24, kashi 38, katana 20, katte 22, kautaa 28, kabin 29

8. **karasu** karada 33, kabin 31, kai 29, kaki (1) 28, kaki (2) 20, kaki 24, kakugai 39, kamakari 34, kameru 30, kanna 29, kanba 20, kanpo 7, kansou 36, kappy 15, kappumen 34, kappuru 36, karahuru 26, karui 22, karamu 12, karuta 20, karute 15, karuteru 1, kasegi 24, kashi 40, kattaa 35, kaze 37, kazoku 39

9. **karada** karasu 12, kaki (1) 19, kamera 23, karabukin 8, kasuma 28, kai 11, kai 18, kana 30, kame 29, kan 11, kai 18, kaibutsu 11, kappe 29, karaku 18, kaminoko 27, kamisama 19, kan 18, karni 14, kappu 17, kappa 11, kappu 15, karaa 18, karaage 31, karukichi 19, karakuiri 11, kamaru 13, karami 4, karamu 8, karappo 19, karashi 18, karasumi 5, karate 15, kashira 5, katusui 12, katsui 14, kattaa 28, katsura 15, kaze 36, karasu 12

10. **karaharu** karamatsu 14, kakato 30, kaki (2) 24, kaki (1) 35, kakke 14, kamera 32, kamisama 30, kanda 18, kantu 29, kappu 34, kara 39, karahuto 26, karai 26, karasu 24, karochin 27, karoushi 26, karui 21, karuta 24, kata 20, katachi 22, kattaa 27, kah 28, kazamuki 28, kazan 21, kazoku 37, katorisenkou 26

11. **karamatsu** karahuru 31, gasa 38, kabi 37, kabin 42, kabuto 34, kaede 40, kaze 33, kaimono 32, kake 33, kaki (2) 36, kaki (1) 29, kame 36, kamera 40, kanti 35, kominari 38, kanada 38, kanyou 39, kappu 28, karai 37, karari 22, karamawari 38, karoni 21, karon 13, karon 30, karasu 39, karu 35, karu 32, karubona 37, karui 35, karute 37, karuwa 12, kasa 38, kasseto 43, kashi 39, kasei 36, katani 23, katsu 34, kaze 36, karasu 12
22. **wakame**
wakare 16, wadachi 7, waido 18, wain 31, waka 10, wakaba 14, wakai 13, wakasa 22, waki16, wakoku , wan 5, wani 14, warutsu 15, watari 4, watashi 36, wara 10

**wakare**
wakame 29, wa 6, wagashi 29, waisyatsu 30, waka 11, wakaba 14, wakadori 28, wakai 21, wakaranai 26, wakaru 17, wakasa 21, wakasagi 6, wakatta 17, wakayama 19, wakka 7, waku 10, wan 13, wankage 10, wappu 2, waribashi 31, waru 18, washig (2) 10, wata 20, wataru 13, watashi 26

23. **warabi**
warai 20, waapu 27, wain 42, waisyatsu 37, wajutsu 27, waka 28, wakaba 26, wakame 39, wakkusu 33, wana 29, waon 19, wara 29, warera 29, warutsu 29, washi (2) 28, washi (1) 30, wataame 21, wani 31

**waraji**
warabi 25, wai 31, wakasigi 9, wakai 35, wakame 38, wakaranai 35, wana 31, wani 34, warabeuta 11, warabukiyane 17, wari 38, waru 3, waribishi 42, warui 34, washi (1) 33, wata 30, watarashii 35, washi 43

24. **yomichi**
yomise 15, yo 17, yogiri 16, yoiko 30, yoka 29, yokaze 23, yomikaki 30, yomimo 25, yominokuni 16, yomu 27, yon 2, yondan 6, yondo 8, yonhon 8, yonko 7, yonk 20, youkan 29, yoru 36, yotto 28

**yomise**
yomichi 30, yobina 28, yoka 35, yokyou 22, yomikiri 26, yomimo 28, yomisute 9, yomiuri 17, yonaka 33, yononaka 34, yoppai 35, yorei 17,yorokobi 36, yosaku 5, yosenabe 34, yosou 22, yotsuba 24, yotto 28, yoko 29, yomi 25, yomikata 35, yomitori 22, yoroi 17, yotsuya 16, yoru 37