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 Zwitserlood11 and Zwitserlood12 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 Hambly3 compared recognition of English words beginning in the same way but ending differently such as difficult and diffident; Slowiaczek, Nusbaum and Pisoni31 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 words5) 6) 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 CVCCV sequence but which, in Tokyo Japanese, have different accentual 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 (Gernsbacher8), Kreuz9) 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 fragment na-, 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 - compyuutaa (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 compyuutaa 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, nagaku, 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
4. kamotsu
3. hokubu
2. hanabi
1. bakuhu
distinguished: e.g. and so on. Two words with the same Roman orthography are implying that 25 subjects rated nagasa than for the example nagasa, 24 subjects rated nagashi as more familiar than the stimulus word. The remainder

9.
10.

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 for the example nagasa, 24 subjects rated nagasa as more familiar than nagashi; 20 subjects rated nagashi as more familiar than nagasa; 20 subjects rated NASA as more familiar than nagasa, implying that 25 subjects rated nagasa as more familiar than NASA; and so on. Two words with the same Roman orthography are distinguished: e.g. washi (1), washi (2).

1. bakuhu bakuchi 19, baasan 35, bagu 37, baiku 40, baioin 38, baka 36, bakku 31, baku 5, bakudan 30, bakuro 37, banana 43, basya 25, batai 41, bataas 37, batto 14 bakuchi baku 26, bacchi 28, bagu 37, baka 36, bakemono 25, bakin 31, bakku 35, bakkunetto 12, baku 4, bakuichiu 18, bakudan 26, basuketto 34, batsu 35, bata 29, batto 32

2. hanabi hanawa 9, hada 31, hage 25, haguki 14, haisha 31, hakkai 5, hakuki 5, hamachi 17, hamaki 9, hamu 30, hanare 12, hanashi 32, hanchou 17, handan 18, hanoru 26, hakachi 34, hanki 5, hanpen 20, hanshi 4, hanten 5, happa 16, happei 16, haru 30, hashika 13, hatena 11, hato 26, hatari 19, hatto 7, hachi 22 hanawa hanabi 36, hadagi 30, hande 30, hana 41, hanami 39, hanamizu 35, hanao 13, hanashi 35, hanauta 30, hanaya 36, hanga 25, hankichi 41, hanki 36, hankouy 34, hanran 27, hansen 26, hantai 12, happa 36, haramaki 20, hari 12, hashi 37, hasu 21, hato 31, hattari 22

3. hokubu hokuro 28, hokutou 35, hotaru 32 hokuro hokubu 16, hoken 20, kokai 6, hokcidou 28, hokke 17, hokkyoku 14, hokurou 9, hon 36, hontou 34, hontou 28, hoshi 32, hotaru 10, hoto 21, hoddokoku 29, houseki 23

4. kamotsu kamome 33, kaisya 40, kaiji 40, kakashi 27, kaki (2) 29, kakko 30, kakkou 23, kaku 29, kamameshi 30, kamera 38, kamikaze 15, kamo 24, kamonabe 10, kamu 35, kan 38, kanitsu 15, kanchi 19, kanpa 23, kanpu 32, kantou 42, kanzume 37, kappu 38, karasu 36, karate 28, karutueru 18, kashi 40, kattai 35, katei 41, katsu 35, kattou 31 kamome kamotsu 12, kami 35, kabu 25, kabuto 13, kachime 25, kaidan 38, kaisya 32, kakashi 22, kakke 12, kakkou 9, kamaboko 33, kamakiri 23, kamisama 40, kamo-resu 14, kamoku 28, kamon 13, kamoniku 14, kan 18, kana 18, kanbotsu 5, kankou 37, kanmuri 13, kanpachi 12, kanso 36, kantan 23, kanten 29, karasu 24, karate 24, kashi 38, katana 20, katte 22, kaitna 28, kabin 29

karasu karada 33, kabin 31, kai 29, kaki (1) 28, kaki (2) 20, kaki 23, kakugou 21, katuru 22, kamera 34, kami 29, kann 8, kanpa 20, kanpo 7, kansou 36, kappum 34, kappuru 36, karahuru 26, karu 22, karume 12, karuta 20, karute 15, karutueru 1, kasegi 24, kashi 40, kattai 35, kaze 37, kazoku 39 karada karasu 12, kaki (1) 19, kamera 23, karabuki 8, kasutera 28, katai 11, katatsumuri 18, katsu 16, katte 14, kai 18, kaibutsu 11, kaki (2) 19, kakkou 18, kaminokke 27, kamisama 19, kan 18, kani 24, kantou 17, kappu 11, kappu 15, karaa 18, karage 31, karkuchi 19, kara 11, karamu 13, karamu 4, karamu 8, karappo 19, karashi 18, karamumi 5, karate 15, kashira 5, kasugai 8, katachi 12, katsute 14, kattai 28, katsura 15, kaze 36, karas 12

karahuru karamatsu 14, kakato 30, kaki (2) 24, kaki (1) 35,kke 14, kamera 32, kamisama 30, kanda 18, kantou 29, kappu 34, kara 35, kararuto 16, karai 26, karasu 24, karochin 27, karoushi 26, karu 21, karuta 24, kata 2 (2) 30, katachi 22, kattai 27, karate 28, kazamuki 28, kazan 21, kazoku 37, katorisenkou 26 karamatsu karahuru 31, gasu 38, kabi 37, kabin 42, karubu 34, kaede 40, kage 33, kamisama 32, kake 33, kaki (2) 36, kaki (1) 29, kame 36, karon 40, kami 38, kaminari 38, kanada 38, kanyou 39, kappa 28, karai 37, karamari 22, karamawari 38, karamu 21, karu 13, kara 30, karasu 39, karate 35, karu 32, karubona 37, karu 35, karuta 37, karut 38, karuwa 12, kasu 38, kasetto 43, kashi 39, kasutera 38, katachi 34, katanomi 36, kattai 33, katsura 34, kaukyo 30 karutureru karuturu 17, kabi 41, karutu 24, kaisou 41, karitsu 40, kaki (2) 32, kaki (1) 40, kakkai 39, kakkou 35, kakagi 38, kame 38, kappu 33, kappu 41, kappur 39, karashi 40, karutu 41, karu 31, karutou 13, karu 37, karuta 36, karute 39, kasi 31, kashi 40, katsuodon 42, kattai 40, katsura 38, kattou 35, kaya 24, kazo 42 karutera karutueru 28, kai 41, kaigara 42, kaimono 32, kaki (2) 39, kakkou 38, kakau 20, kami 38, kamo 38, kamotsu 38, karaage 42, karasu 42, karate 36, karuchia 41, karuashahokkou 38, karugamo 37, karu 32, karusih 35, karu 36, karume 39, karumen 34, karusa 43, karusuyumu 43, karuta 38, karutatorui 35, karute 40, kashi 41, kake 1 (1) 32, katachi 36, katai 36, kataka 41, katana 33, kateri 39, katsu 40, kazoe 1 (1) 39, kazo 41, kazumo 36 kasegi kasetu 25, kaitten 35, kabo 30, kubuto 16, kajitsu 33, kaki (1) 33, kame 26, kamera 37, kami 35, kan 37, kanban 25, kann 9, kanro 13, kappu 19, karasu 30, kariin 24, karuta 28, kasa 34, kasei 25, kaseki 13, kase (1) 22, kase 23, kasetto 39, kashi 40, kasu 12, kasumi 13, kasumissou 29, kasutanetti 26, kata 2 (2) 32, katatsumuri 30, kato 12, kattai 37, kaze 1 (1) 36 kasetsu kasegi 24, kachki 26, kai 2 (2) 30, kaki 1 (2) 27, kaki 32, kamera 38, kamisama 33, kappu 27, karasu 28, kasetto 37, kasa 27, kasu 15, kasukabe 22, kasumissou 35, kasuri 17.
14. namida
13. nagasa
12. mokuba
11. moguri
10. maguchi
9. kokugi
8. naihu 8, nameko 18, nami 17, nana 2, nanda 8, nanzan 6, naru 8, nasubi 11, nasu 18
7. namiki
6. namikage 28, namikura 30, namikomu 14, namikoku 31, namikusa 14, namida 41, namikuzu 30, nami 18, nami 18, namidame 24, namikaze 18, namimichi 23, namima 28, naminori 14, nanten 20, nashi 33, natsu 37, natsumikan 34, natto 39
5. nimotsu
4. naito 22, naihu 11, nihon 34, niki 32, nikko 32, nikkou 32, niku 36, ninningu 36, ninnin 36, ninnyou 27, nizu 31, nishi 12, nippou 12, ninnja 21
3. nimo
2. nohara 33, nokogiri 29, nomi 21, nomi 3, nomimono 41, nomu 36, nohitei 18, nori 35, nonki 37, noppou 29, nouchou 37, nouka 30, noutou 34
1. nomiya
10. morikiri 24, mabuta 45, macchi 42, mada 40, maguma 35, makura 43, man 39, manto 39, matsu 44, matsuou 35, matsuura 34, matushita 18, mato 14
11. moguri
9. mogura 31, modomu 37, mugosa 14, moji 41, mokkin 13, mokkoku 3, mokkou 7, mou 16, moomoo 24, mokuyoubi 37, mokuzai 16, mokudoku 11, mokugyo 18, mokuroku 9, mokusei 20, mokutan 18, mokujou 30, mottou 37, motsunabe 32, motsuyaki 30
12. mokuba
13. nagasa
14. namida
20. wakasagi 31, wakatta 33, wakayama 21, waki 22, wakka 9, wakkanai 17, wakusei 27, washi (1) 25, washi (2) 14, wasyoku 40, wata 25, watagashi 32, watashi 37

22. wakame 29, wakare 16, wadachi 7, waido 18, wain 31, waka 10, wakaba 14, wakai 13, wakasa 22, waki16, wakoku 5, wani 14, warutsu 15, wataru 4, washi 36, wara 10

wakare 29, wa6, wagashi 29, wasyatsu 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, wanege 10, wappa 2, waribashi 31, waru 18, washi (2) 10, wata 20, wataru 13, watashi 26

23. warabi 29, waapu 27, wain 42, wasyatsu 37, wajutsu 27, wak 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 25, wan 31, waragutsu 9, wakai 35, wakame 38, wakaranai 35, wananai 31, wani 34, warakeuta 11, warakutigi 17, warai 38, waru 3, waribashi 42, warui 34, washi (1) 33, wata 30, watagashi 35, watashi 43

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

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