Shoju CHIBA (Reitaku University, Japan) e-mail: schiba@reitaku-u.ac.jp

Theory and Practice of Enriching a Word List

A Case Study of Building a Student Glossary of Finnish for Japanese Learners

RDHum 2019 @ University of Oulu 2019-08-14

Outline of the Presentation

- Word Lists? Glossaries?
  - Existing word list/glossaries of Finnish
  - Vocabulary features of Finnish language and why the word list is needed
- Building a word list for Japanese learners
- Evaluating the word list: The outlines of the current project
- Enriching a Word List with Word2Vec Technology
- Summary

Word List for Learners: an Introduction

Existing Word Lists of Finnish Language

- Only few found for students
- There are quite a few published tourist phrase books of Finnish, even in Japanese.

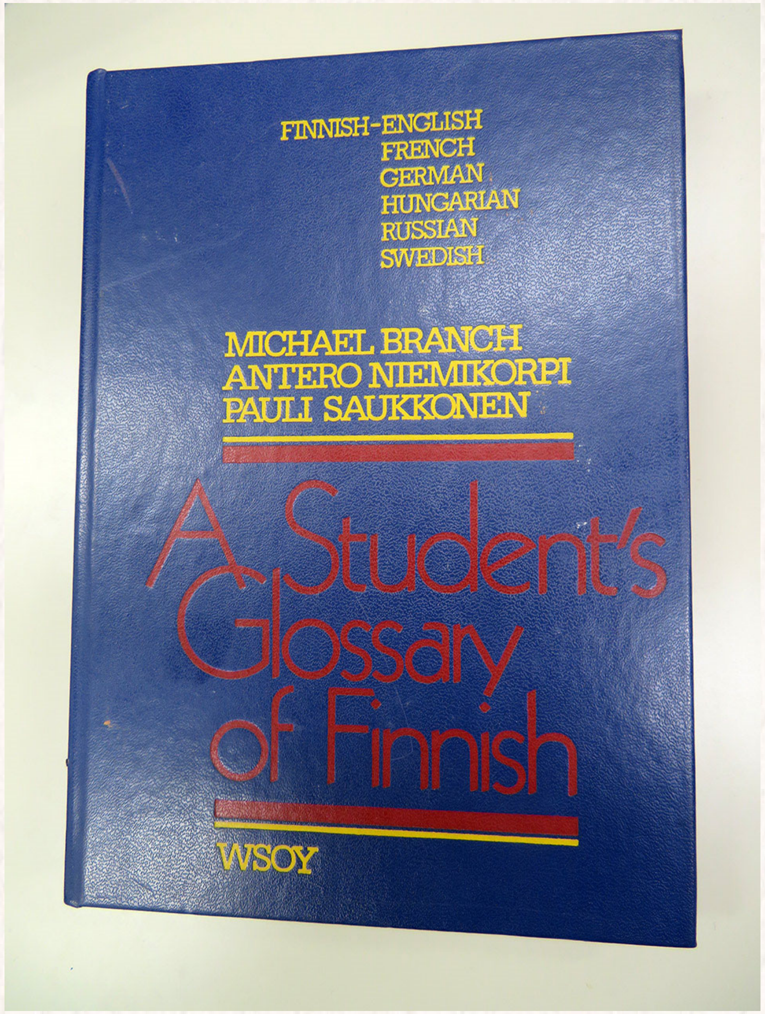
- Branch, Michael, Antero Niemikorpi & Pauli Saukkonen (1980) *A Student's Glossary of Finnish*. Porvoo: WSOY.

As for Japanese word lists:

- Takashi Ogishima (1990) *1500 Basic Words of Finnish*. Tokyo: Daigaku Shorin. (**荻島崇 『フィンランド語基礎1500語』大学書林**)
- Kingo Yoshida (2019) Passport Basic Dictionary of Finnish. (A Finnish-Japansese dictionary with 5,700 head words) Tokyo: Hakusuisha. (**吉田欣吾『パスポート初級フィンランド語辞典』白水社**)
  - Appendix 2 (Theme-based glossaries of Finnish, around 1,000 words)

Branch et al. (1980)

- Based on the Oulu Corpus (155,000 running words)
- Picks up the words with ten or more times in the Corpus (about 2,000 words) and arranged with their rank codes (frequencies)
- An important outcome of the dawning age of corpus-based Finnish studes



”Requirements of Student Glossary Worth Consulting”

Editorial board of the English word list *New JACET8000* (2016:9)
*JACET = The Japan Association of College English Teachers

- Relevance = To what kind of learners the word list will be presented
- Verifiability = Whether the word list is reproducible
- The importance of scientific compilation of the word list, where the production process of it is well examined

Relevance matters

Nation (2004) compares the most frequent 3,000 words taken from BNC with General Service List (West 1953) and Academic Word List (Coxhead 2000)

- Distribution of the vocabulary of BNC is different
  - “it is not appropriate to use these lists unchanged as the basis for syllabus design for learners of English as a second or foreign language in primary or secondary school systems.” (Nation 2004:3)
- BNC lists reflected the adult, British, formal nature of BNC, which cause a serious mismatch if used in primary or secondary school systems.
- The vocabulary referred to cannot be decided without considering the teaching materials and the examinations used in the syllabi of the school systems concerned (New JACET8000:9)

Consistency of frequency and distribution

Scott & Tribble (2006) *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

- Scott & Tribble (2006:29-30): In addition to frequency, the consistency contributes to the examination of the lexical items
  - Characteristics of the lexical items
  - Uneven distribution of the items among different genres
  - Special nature of the text itself where the items emerge
- Vocabulary features serveyed via large corpora can be useful to check the feasibility of the word list

A note on the lexical characteristics of Finnish language

- Vocabulary proper to Finnish (or Finnic, Finno-Ugric): analogy or similar reasoning via mother tongue doesn't work
- The number of the basic words used in everyday life well surpasses 4,000 words (Leena Silfverberg, p. c.)
  - Morphological complexities impact this large number of basic words heavily
  - Acquaintance with morphological information (derivational suffixes, for example) helps learners to understand the lexical network of the vocabulary

Demands of the materials for the lexical study of Finnish language

There ARE quite a few dictionaries of Finnish language available: the list of the dictionaries is found e.g. in my home page http://www.tsibale.com/sanakirjat.html

- There are too few word lists for reference
  - Ogishima (1990) 1,500 word list for Japanese learners
  - Branch et al. (1980) 1,899 word list in 6 languages
- Few researches handling the vocabulary of Finnish language for everyday/student use
- No word list compiled according to the different study level (cf. JACET8000); Branch et al. (1980) classifies the words according to the frequency rank only.
- In CEFR-based teaching materials, the selection of the vocabulary is mainly left to the compilers' subjective belief (cf. Council of Europe 2001, § 6.4.7.3)

Building a word list for Japanese learners

Aims of compiling a student glossary for Japanese learners

- For the convenience of learners
  - To get acquainted with Finnish vocabulary
  - Acquiring a certain amount of basic words is essential
- Word list for beginners: Browne’s New General Service List (Browne 2014) might be the optimal sample product
  - The situation has many points of similarity to the background of compiling Branch et al. (1980)
- Special attention to the environments of Japanese learners of Finnish
  - Poor learning environment
  - Remoteness from Finland
  - Difficulty of keeping motivation cf. demand of English for high school/university entrance examinations

“A glossary of Finnish for Reitaku Students” (Tsuchiya 2019)

- 12 students participated in the special seminar to build the original word list and learn Finnish vocabulary in spring term 2018
- 11 categories (themes) are set beforehand
- In each category students picked up the word which they want to learn
  - Listed the word in GoogleDrive
- The teacher checked and complemented the list
- Finnish native students checked the list
- Participants shared the list with an online flashcard application (Quizlet) and studied the words before the seminar

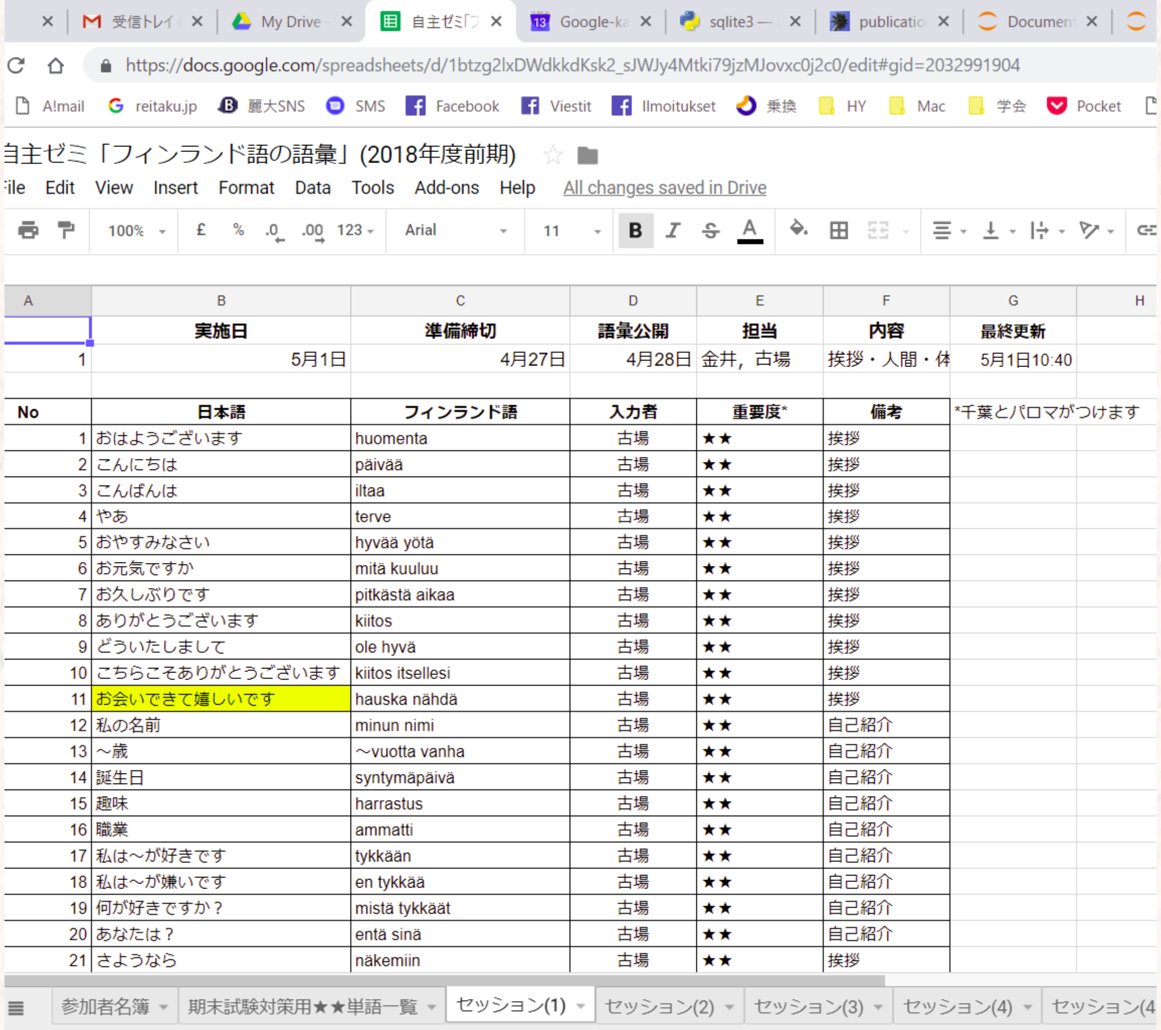
Schedule for preparation and lexical study

- 12 students participated in the special seminar to build the original word list and learn Finnish vocabulary with it in spring term 2018
- In total, we hold 11 sessions and made a 2,080 word list

No	Date	Category
1	May 1st	Greetings, human relationship, body
2	May 8th	City, country, transportation
3	May 15th	Food
4	May 22nd	Clothes, sports, culture
5	May 29th	Action, activity
6	June 5th	Color, feeling, adjectives with pair
7	June 19th	Room, interiors
8	June 26th	Weather, season, time
9	July 3rd	School and school/University life, social life
10	July 10th	Nature
11	July 17th	Time and place

Input format @ GoogleDrive

- In total, we hold 11 sessions and made a 2,080 word list
- Teachers and native students marked star(s) for essential words, the number of which was currently 522 words



https://docs.google.com/spreadsheets/d/1btzg2lxDWdkkdKsk2_sJWJy4Mtki79jzMJovxc0j2c0/edit#gid=2032991904

自主ゼミ「フィンランド語の語彙」(2018年度前期)

All changes saved in Drive

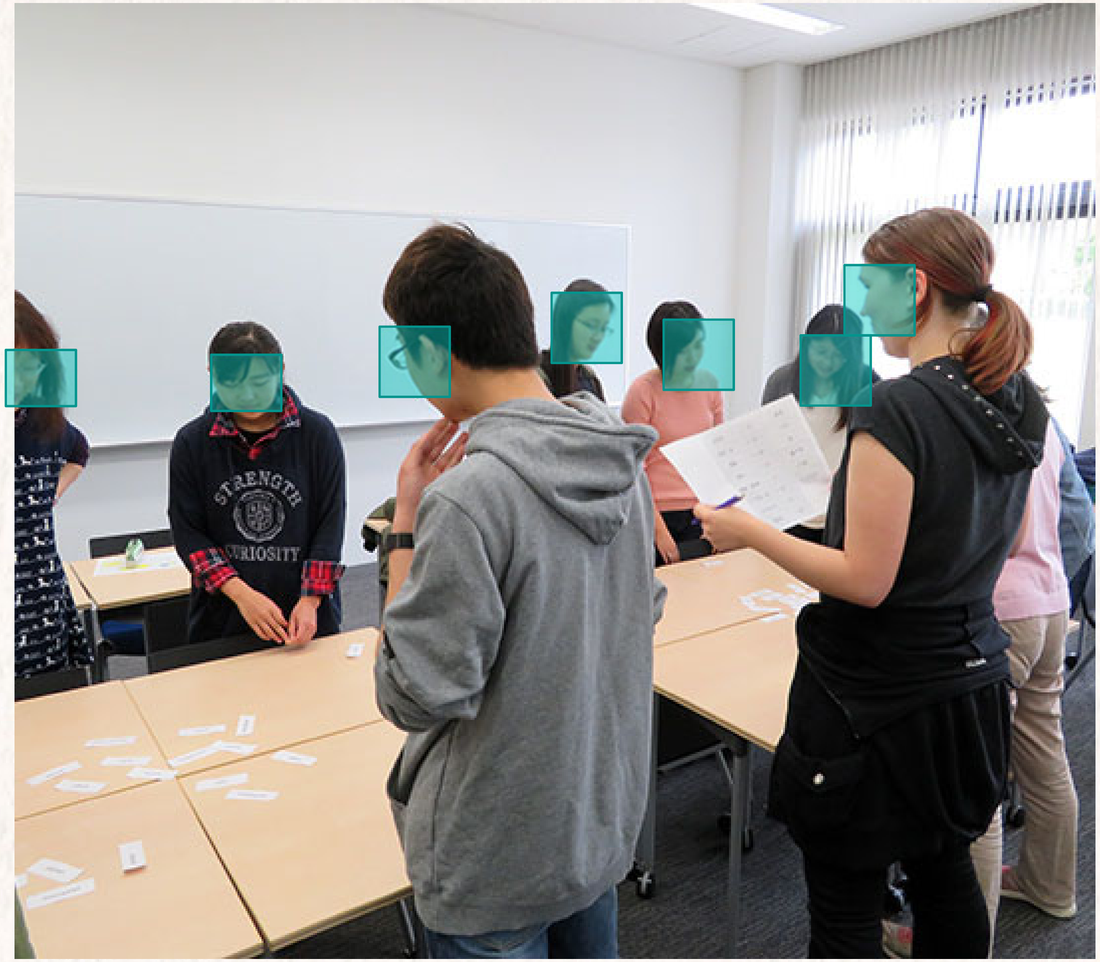
	実施日	準備締切	語彙公開	担当	内容	最終更新
1	5月1日	4月27日	4月28日	金井，古場	挨拶・人間・体	5月1日10:40

No	日本語	フィンランド語	入力者	重要度*	備考	*千葉とパロマがつけます
1	おはようございます	huomenta	古場	★★	挨拶	
2	こんにちは	päivää	古場	★★	挨拶	
3	こんばんは	iltaa	古場	★★	挨拶	
4	やあ	terve	古場	★★	挨拶	
5	おやすみなさい	hyvää yötä	古場	★★	挨拶	
6	お元気ですか	mitä kuuluu	古場	★★	挨拶	
7	お久しぶりです	pitkästä aikaa	古場	★★	挨拶	
8	ありがとうございます	kiitos	古場	★★	挨拶	
9	どういたしまして	ole hyvä	古場	★★	挨拶	
10	こちらこそありがとうございます	kiitos itsellesi	古場	★★	挨拶	
11	お会いできて嬉しいです	hauska nähdä	古場	★★	挨拶	
12	私の名前	minun nimi	古場	★★	自己紹介	
13	～歳	～vuotta vanha	古場	★★	自己紹介	
14	誕生日	syntymäpäivä	古場	★★	自己紹介	
15	趣味	harrastus	古場	★★	自己紹介	
16	職業	ammatti	古場	★★	自己紹介	
17	私は～が好きです	tykkään	古場	★★	自己紹介	
18	私は～が嫌いです	en tykkää	古場	★★	自己紹介	
19	何が好きですか？	mistä tykkäät	古場	★★	自己紹介	
20	あなたは？	entä sinä	古場	★★	自己紹介	
21	さようなら	näkemiin	古場	★★	挨拶	

参加者名簿 | 期末試験対策用★★単語一覧 | セッション(1) | セッション(2) | セッション(3) | セッション(4) | セッション(4

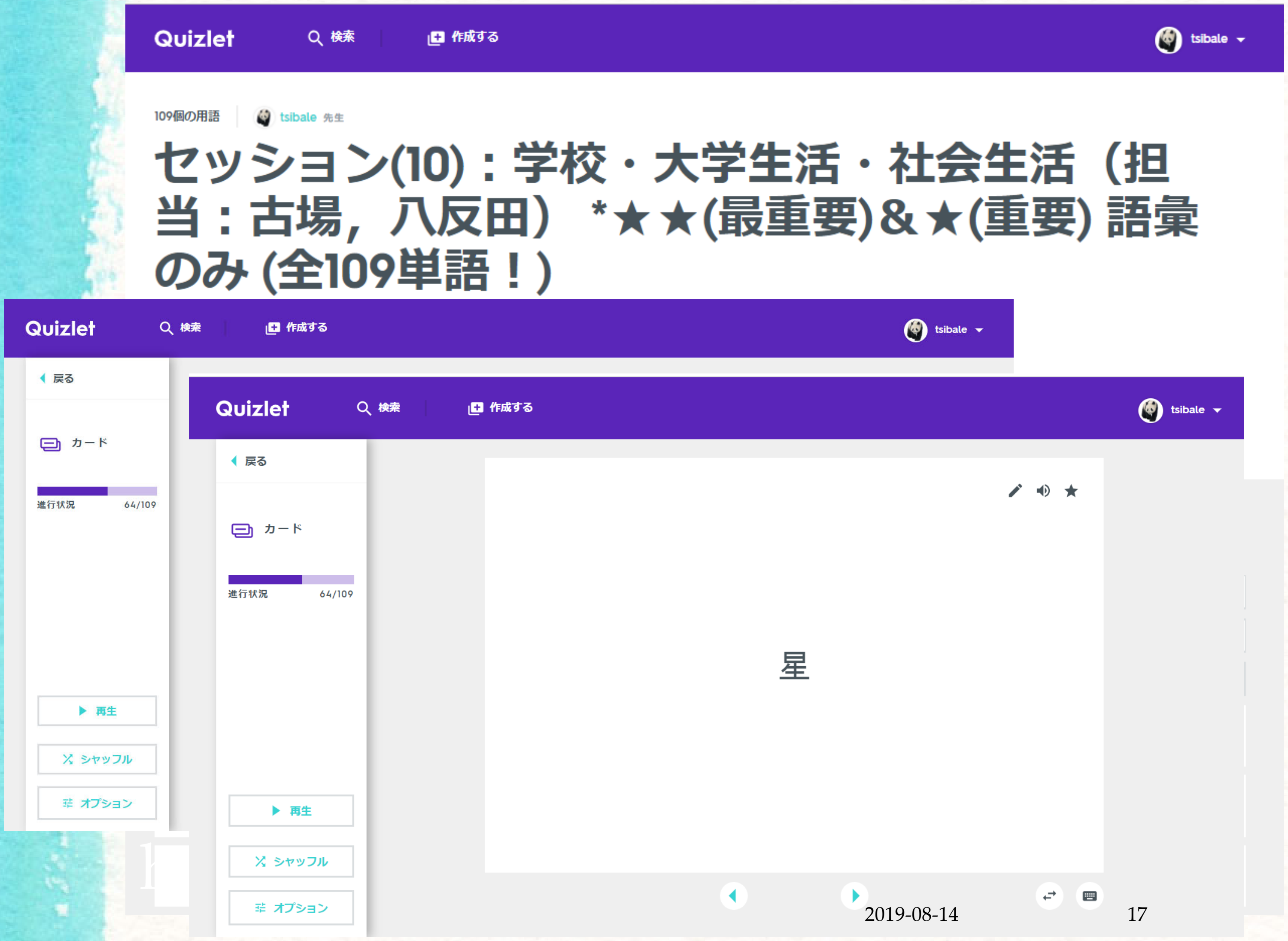
Study scene with the word list

The style of **karuta**, traditional Japanese playing cards, was adopted. Native students helped to pronounce the flash cards



How to study with Quizlet

https://quizlet.com



109個の用語 tsibale 先生

セッション(10)：学校・大学生活・社会生活（担当：古場，八反田） *★★(最重要)＆★(重要) 語彙のみ (全109単語！)

進行状況 64/109

星

Scores of the participants

- Studying around 100 words in a week with Quizlet seemed a hard task for many of students…

Score (cumulated)	Session1 Full list	Session1 Selected	Session2 Full list	Session2 Selected	Session3 Full list	Session3 Selected	Session4 Full list	Session4 Selected	Session 4.5 (follow up)	Session5 Full list	Session5 Selected	Session6 Full list	Session6 Selected
4.5		1.5	0.5			0.5		0.5	0.5		0.5		0.5
11	1	3.5	1	1	1	0.5	1		0.5	1		0.5	
17.5		2		2.5		3	1.5	2	2.5		2		2
0													
3.5		2		0.5								0.5	0.5
0													
0													
11			1.5			2		1	1	1		1	3.5
1						1							
16	1	2.5		2.5		1.5		2.5	2.5		1		2.5

Publishing a refined word list

is the ultimate purpose of the study

- There are few word lists for Japanese learners of Finnish, which have the wide coverage of the topics and consider the difficulties of the lexical items
- The list is made ad hoc according to the choice of the student (in assistance of the teacher and native speakers), so the evaluation (and refinement) of the list is inevitable
- No comparable data: There are not sufficient lexical data on the vocabulary of Finnish used in daily/University life
- No spoken corpus of Finnish language with relevant size/registers

Project 1: Questionair method

Currently carrying out the project

- Evaluation of the importance/relevance of the words by distributing the word list to:
  - Native Finnish student studying in Japan
  - Native Finnish teachers teaching in Japan
  - Native Finnish teachers teaching in Finland
  - Japanese Finnish teachers teaching in Japan
- Evaluation with 4 grade (not important – very important) + options "not sure"
- Calculating kappa coefficient
- Analyzing with multiple comparison method

Project 2: Corpus-based method

Currently evaluating and analyzing the word list

- Evaluating the “importance” of words using the corpus data in 3 viewpoints
  - Frequency information
  - TF-IDF (“weight” of the word)
    - TF: Term Frequency
    - IDF: Inverse Document Frequency
  - Resemblance indices of Latent Semantic Analysis, (LSA): by extracting the related words from the corpus, compare the frequency and TF-IDF of the items of the word list with the related words

Latent Semantic Analysis (LSA) and Word2Vec

Distributional Hypothesis (DH)

as a theory of linguistic semantics

- “a word is characterized by the company it keeps” (Firth 1968)
- “If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution” (Harris 1954)

Distributional Hypothesis in detail

Evert & Lenci (2009)

- At least certain aspects of the meaning of lexical expressions **depend on their distributional properties in the linguistic contexts**
- The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear (Harris 1954)
- The relationship between word meaning and word usage in contexts matters!

Distributional Semantic Models (DSMs)

Fabre & Lenci (2015)

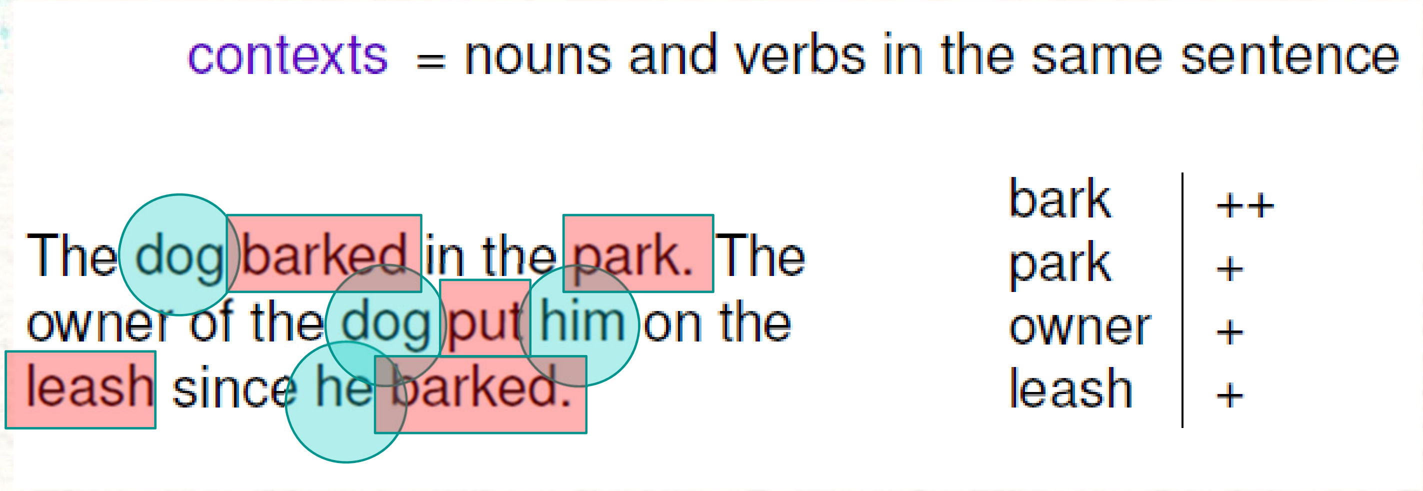
- DSM is a model to extract from the corpus data the meaning which the context shows
- By processing context information statistically, the model calculated the vector information and express the semantics with vectors

How to vectorize distributional information

- Count the word in a context
- Calculate the vector by counting the frequency in a context
  - This assumes the word with similar meaning shows similar vector
  - distributional similarity ≒ semantic similarity
- Visualize the meaning differences with vector difference

How to count context and frequency

Example taken from Evert & Lenci (2009)



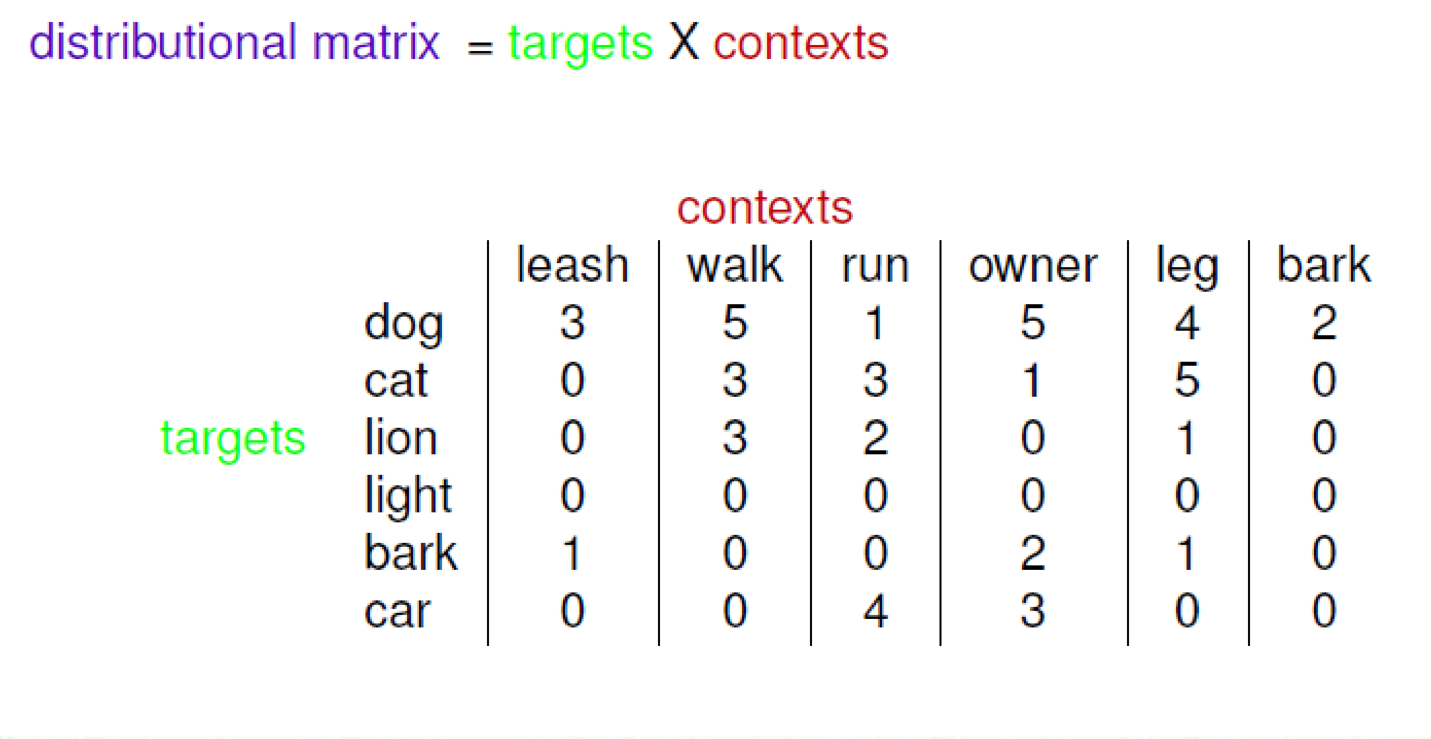
contexts = nouns and verbs in the same sentence

The dog barked in the park. The owner of the dog put him on the leash since he barked.

bark	++
park	+
owner	+
leash	+

Target-wise counti of the context

Example taken from Evert & Lenci (2009)

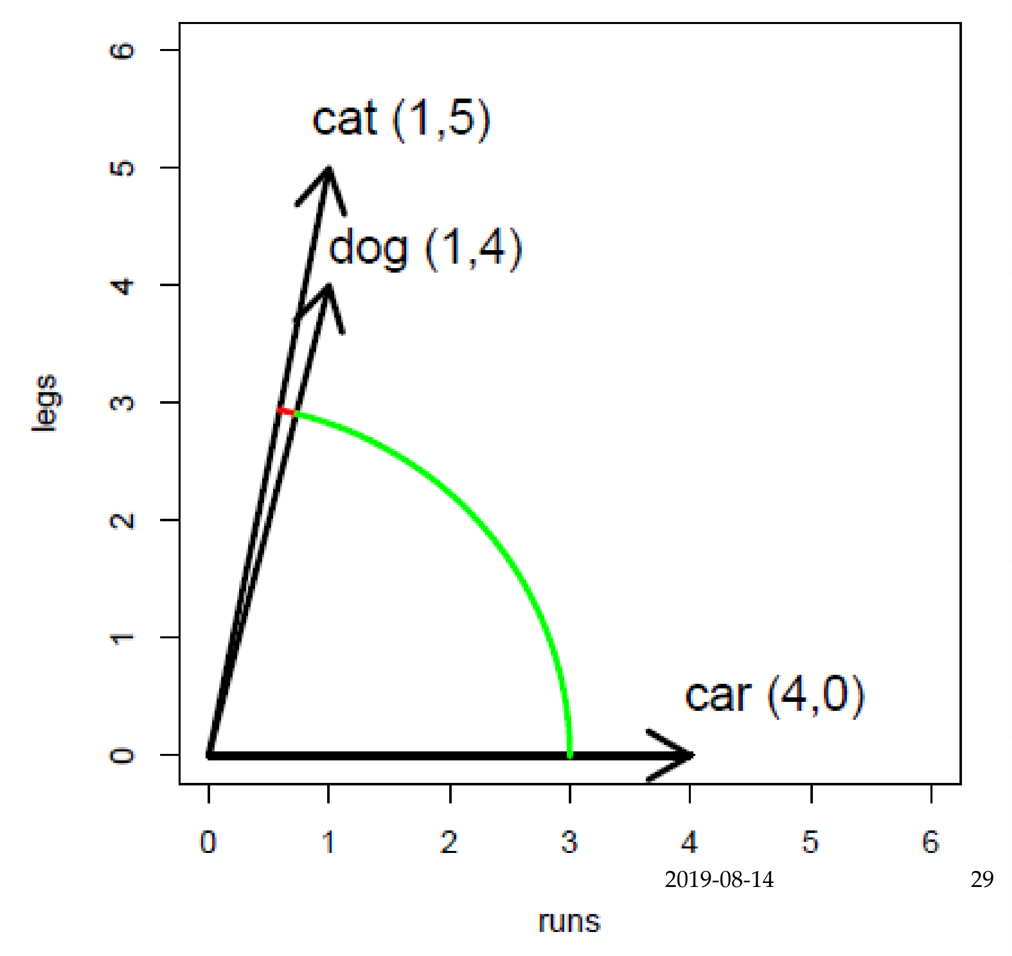


distributional matrix = targets X contexts

		contexts					
		leash	walk	run	owner	leg	bark
	dog	3	5	1	5	4	2
	cat	0	3	3	1	5	0
targets	lion	0	3	2	0	1	0
	light	0	0	0	0	0	0
	bark	1	0	0	2	1	0
	car	0	0	4	3	0	0

Vectorizing the context

- Difference of the degree of angles express the semantic difference of the words



Significance of DH

and its application to natural language processing and lexical study

- WEAK DH assumes a certain connection between context and mieaning
- Postulating this weak version of DH, we can apply the method to
  - Disambiguation
  - Development of automatic response (unsupervised) and other recommendation system
  - Lexicography and other lexical analysis
    - Distributionally similar = semantically similar
    - Extraction of related vocabulary
- Current study is in line with this version of DH.

Significance of DH (continued)

and its application to cognitive science

- Strong DH again assumes the strong tie between semantic expressions and the cognitive model
  - Distribution of the words in context has close relationship with the semantics they express, thus explain how the semantics are structured
  - Human semantic behavior (judgement of semantic similarity) can be explained using the distributional nature of words
  - Cognitive model, conceptual model (AI, deep learning)
    - Not only learning of words, but
    - Learning of the semantic categories

Defining Distributional Semantic Model (DSM)

- DSM is a set of < *T*, *C*, *R*, *W*, *M*, *d*, *S* >
  - *T* = target word (which DSM gives context)
  - *C* = context where T appears (defining window-size and weighting is important)
  - *R* = relationship between T and C (sentence-internal, paragraph-internal, turn-internal; syntactic/lexical dependency relationship)
  - *W* = scheme of the weight of context
  - M = distributional matrix, T×C
  - *d* = function of dimension reductiond : M → M′
  - *S* = distance in the vector space M′

Types of DSM and their parameter settings

- To vectorize the context information, you need decide
  - Context type
  - Weighting scheme
  - Method to analyze the similarity
- There are numerous DSM types proposed, and the settings of the parameters above varies
  - The results might be completely different
  - This study utilizes Word2Vec to evaluate the word list

word2vec

Mikolov et al. (2013)

- Mikolov et al. (2013) deviced an efficient neural network model for vectoralization of the context information
- The model has two distinct types of calculation measure: high-speed **CBoW** and more accurate **Skip-gram**
- Well known for its high precision of vector expression of semantic relationship: thus the model is applied to various natural language processing tasks like
  - Machine translation
  - Named-entity recognition
  - Sentiment analysis

Application of Word2Vec

Studies I am pursuing currently using Vector analysis

- Case 1: Analyzing the characteristics of the corpus of learner Finnish (Chiba 2016, 2017)
- Case 2 : Evaluating the word list of Finnish language (current study)

Evaluation process with vectorized data of this study

- Using the Word2Vec data of a large-scale corpus of Finnish (*Finnish Internet Parsebank*, comprising 1,500 million words), extract the list of the related words of each lexical item of the word list, with

- Compare the frequency and TF-IDF of the lexical items of the word list of the original glossary and the one suggested by Word2Vec

- Complement the word list by adding the relevant related words with high frequency and great TF-IDF

Sample Analysis: Enriching a word list with Word2Vec technology

A sample case

- *Finnish Internet Parsebank* by Turku BioNLP Group, which contains 1.5 billion tokens.
  - word2vec data is also provided by the same team. (In detail, see Kanerva et al. 2014)
  - Home page URL: http://bionlp.utu.fi/finnish-internet-parsebank.html
- Lemma-based analysis is not always the best
  - Jantunen et al. (2013) points out that words shows the different profile with different word-form frequency

Tokenized data in comparison

Nearest neighbours of ’hieno’

- 'mahtava', 0.8341175317764282
- 'upea', 0.7793271541595459
- 'loistava', 0.7552661895751953
- 'mainio', 0.7503886222839355
- 'hauska', 0.7275123596191406
- 'kaunis', 0.7222684621810913
- 'nätti', 0.7192226052284241
- 'mielenkiintoinen', 0.7180345058441162
- 'ihana', 0.6754920482635498
- 'erikoinen', 0.6725770235061646

Nearest neighbours of ’hienoa’

- 'mahtavaa', 0.8150333762168884
- 'upeaa', 0.749343752861023
- 'upeata', 0.688971996307373
- 'loistavaa', 0.6719725131988525
- 'mielenkiintoista', 0.6703432202339172
- 'ilahduttavaa', 0.6279211044311523
- 'jännittävää', 0.6108795404434204
- 'liikuttavaa', 0.6075063943862915
- 'kiva', 0.6073362827301025
- 'surullista', 0.598127961158752

Different pre-processing

Tokenized corpus

- 'mahtava', 0.8341175317764282
- 'upea', 0.7793271541595459
- 'loistava', 0.7552661895751953
- 'mainio', 0.7503886222839355
- 'hauska', 0.7275123596191406
- 'kaunis', 0.7222684621810913
- 'nätti', 0.7192226052284241
- 'mielenkiintoinen', 0.71803450584
- 'ihana', 0.6754920482635498
- 'erikoinen', 0.6725770235061646

Lemmatized corpus

- 'hie|noja', 0.7627484798431396
- 'mahtava', 0.7021496295928955
- 'upea', 0.6931349635124207
- 'loistava', 0.6724129915237427
- 'nätti', 0.6263025999069214
- 'kiva', 0.6128658056259155
- 'kaunis', 0.6065711975097656
- 'wau', 0.5999494194984436
- 'loistokas', 0.5988346934318542
- 'hieno|jakki', 0.596953809261322

The state of arts of the current study

- Project 1: currently undertaken. The consistency (or discrepancy) of the judgements of Finnish native students/teachers and Japanese teachers is anticipated.
- Project 2: The stared items (important, highly important, 522 words) are roughly analyzed. The outcome is as follows:
  - We find the semantically similar words with higher frequency in about 60% of the lexical items of the word list.
  - Within the similar words we found via Word2Vec evaluations, about 100 words are not found in the original word list.
  - The findings above show that the evaluation based on the vector information works fine.
- Project 2: Evaluation of the remaining lexical items are currently undertaken.
- Final outcome: Refinement of the word list and publication of it: First publication of the word list is scheduled in October.

Future perspectives

- The current evaluation is being undertaken using the Web corpus (*Finnish Internet Parsebank*). In subsequent analysis we will utilize the use of the corpus of more colloquial nature (*Suomi24*, appr. 2400 thousand words). *Suomi 24* will suffice because it contains different genre metadata.
- Our pilot study shows that no-lemma unit-based analysis of Word2Vec is essential and even inevitable to evaluate the word list.
  - The corpus with different morphosyntactic/semantic/meta-linguistic annotations are valuable in evaluating the vectorized semantic behaviors of words
  - The study to improve the accuracy of the evaluation (i.e. to detect the relevant semantic neighbors more efficiently) comes in place

Thank you for your attention!

The first draft of refined word list will be published in October. Details will be announced in my home page: http://www.tsibale.com

References

- Branch, Michael, Antero Niemikorpi & Pauli Saukkonen (1980) *A Student's Glossary of Finnish*. Porvoo: WSOY.
- Browne, Charles (2014) "A new general service list: The better mousetrap we've been looking for?" *Vocabulary Learning and Instruction* 3: 1-10.
- Harris, Zellig S. (1954) "Distributional Structure," *Word* 10: 146-162.
- Jantunen, Jarmo Harri & Sisko Brunni (2013) "Morphology, lexical priming and second language acquisition A corpus-study on learner Finnish," in granger, Sylviane, Gaëtanelle Gilquin & Fanny Meunier (eds.) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. (Corpora and Language in Use – Proceedings 1) Louvain-la-Neuve: Presses universitaires de Louvain, pp. 235-245.
- The Japan Association of College English Teachers (2016) *A New JACET List of 8000 Basic Words*. Tokyo: Kirihara Shoten.
- Kanerva, Jenna, Juhani Luotolahti, Veronika Laippala & Ginter Filip (2014) "Syntactic n-gram collection from a large-scale corpus of Internet Finnish," in *Proceedings of the Sixth International Conference Baltic HLT*.
- Mikolov, Tomas, Kai Chen, Greg Corrado, & Jeffrey Dean (2013) "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at International International Conference on Learning Representations* (ICLRI), 2013.

References (continued)

- Nation, I. S. P. (2004) “A study of the most frequent word families in the British National Corpus,” In Boaards, P. & B. Laufer (eds.) *Vocabulary in a Second Language*. Amsterdam: John Benjamins. pp. 3-13.
- Ogishima, Takashi (1990) *1500 Basic Words of Finnish*. (In Japanese) Tokyo: Daigaku Shorin.
- Scott, Mike & Christopher Tribble (2006) Textual Patterns: Key Words and Corpus Analysis in Language Education. Amsterdam: John Benjamins.
- Teubert, Wolfgang (2004) “Language and corpus linguistics,” in Halliday, M. A. K., Wolfgang Teubert, Colin Yallop & Anna Čermáková (eds.) *Lexicography and Corpus Linguistics: An Introduction*. London: Continuum, pp. 73-112.
- Tsuchiya, Shieri (2019) “A glossary of Finnish for Reitaku Students,” in *Reitaku Information Technology Seminar Papers 2018*. Kashiwa: Reitaku University, pp. 19-22. http://www.cs.reitaku-u.ac.jp/msemi/grad-presen/2019/201901-r.pdf
- Word2vec Homepage: https://code.google.com/archive/p/word2vec/
- Yamauchi, Hiroyuki (ed.) (2013) *Jissen Nihongo Kyôiku Standard*. (In Japanese) Tokyo: Hituzi Shobô.
- Yoshida, Kingo (2019) *Passport Basic Dictionary of Finnish.* (A Finnish-Japansese dictionary with 5,700 head words) Tokyo: Hakusuisha.