TransPhoner: Automated Mnemonic Keyword Generation

Manolis Savva, Angel X. Chang, Christopher D. Manning and **Pat Hanrahan**
Computer Science Department, Stanford University
{msavva, angelx, manning, hanrahan}@cs.stanford.edu

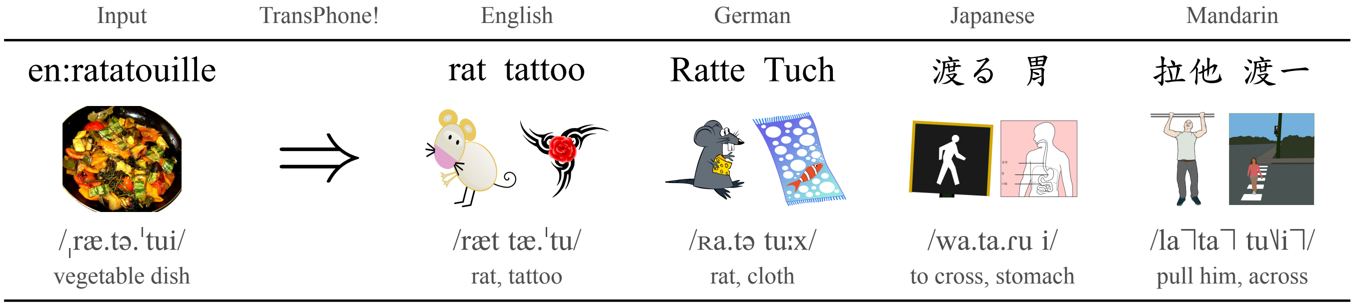


Figure 1. TransPhoner: a system that given terms in one language generates phonetically similar, highly imageable keywords. These keywords can aid learning of foreign language vocabulary and can serve as memorable pronunciation guides (click keywords for audio).

ABSTRACT

We present *TransPhoner*: a system that generates keywords for a variety of scenarios including vocabulary learning, phonetic transliteration, and creative word plays. We select effective keywords by considering phonetic, orthographic and semantic word similarity, and word concept imageability. We show that keywords provided by TransPhoner improve learner performance in an online vocabulary learning study, with the improvement being more pronounced for harder words. Participants rated TransPhoner keywords as more helpful than a random keyword baseline, and almost as helpful as manually selected keywords. Comments also indicated higher engagement in the learning task, and more desire to continue learning. We demonstrate additional applications to tasks such as pure phonetic transliteration, generation of mnemonics for complex vocabulary, and topic-based transformation of song lyrics.

Author Keywords

Mnemonic keywords

ACM Classification Keywords

H.5.m Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Languages; Human Factors; Design.

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.
Copyright © 2014 ACM 978-1-4503-2473-1/14/04..$15.00.
http://dx.doi.org/10.1145/2556288.2556985

INTRODUCTION

> *"The limits of my language means the limits of my world."*
> — Ludwig Wittgenstein

Learning vocabulary is a hard and laborious task, whether memorizing foreign language words or mastering complex terminology. Mnemonic keywords are a learning tool that can be applied to vocabulary learning and other tasks. For example, the keywords and images in Figure 1 can facilitate learning and recall of the English word ratatouille. We present *TransPhoner*: a cross-lingual system that given words in one language, suggests phonetically and semantically relevant keywords in the same or another language.

Prior work has shown that keyword association can improve memorization and pronunciation of foreign vocabulary [2, 13]. However, to the best of our knowledge, there are no existing methods for generating such keywords automatically. To use mnemonic methods, teachers and learners expend considerable effort in manually generating mnemonic material.

Our main contribution is a keyword generation system, with design principles grounded in results from cognitive psychology. To empirically evaluate the effectiveness of TransPhoner keywords we used them for the concrete application of foreign language vocabulary learning. In a human participant study, we found that TransPhoner keywords improve short-term learning performance significantly, with the effect being stronger for harder words. Study participants rated TransPhoner keywords higher on a helpfulness scale compared to a baseline random keyword condition. Finally, we present additional applications of TransPhoner to illustrate the variety of scenarios where keyword generation can be beneficial.

RELATED WORK

Our research is enabled by prior work in psycholinguistics and natural language processing, and related to computational systems for assisting language learning.

Natural Language Processing

Recent NLP research has automatically re-spelled English words to create spellings that clarify pronunciation [18]. The focus was on using phonetically unambiguous English syllables to correct for phonetic inconsistencies in English orthography (spelling). In contrast, we jointly consider phonetic similarity, imageability and semantics to find effective mnemonic keywords in any target language.

Mapping the script of a word from one language to a conventional form in another while preserving pronunciation is the goal of machine transliteration systems—most commonly used for proper nouns (names of people, places and organizations). A recent survey is provided by Karimi et al. [21]. Transliteration approaches differ from our task in that they aim to retrieve typical, commonly accepted transliterations from one language to another. They do not aim to output strings of semantically meaningful words in the target language, nor are they concerned with memorability or learning of a source language term.

Mnemonic Keyword Strategy

Psychologists have long known that mnemonics are a powerful learning strategy [30]. In the specific scenario of vocabulary learning, mnemonic keywords are words presented in addition to the foreign word and its translation. Mnemonic keywords have been extensively studied and are an established learning strategy amongst educators [9, 34]. Fig 2 provides an example. The keyword "cook" is associated with the German for kitchen: "Küche". Typically context is provided in the form of a sentence or image linking keyword with foreign word: "Imagine your kitchen and a *cook* in it".

Keywords are selected to be phonetically similar to the foreign word and highly *imageable*. Imageability is defined as the ease with which a word gives rise to a sensory mental image [31]. High imageability keywords make it easier to visualize interactions between the keyword, foreign word form and underlying concept.

The pioneering work of Atkinson et al [2] showed the effectiveness of keywords for improving vocabulary learning and was followed by studies confirming recognition and recall improvements in a variety of conditions [13, 15, 32]. However, as far as we are aware, our work is the first exploring computational methods for generating mnemonic keywords.

Computer-Assisted Language Learning

The application of technology to language learning has a long history of prior work [1, 22]. We focus on the automated generation of keywords that can help to link a word form—either visual or auditory—with a mental image, an important prerequisite before learning can occur.

Some CALL systems have addressed pronunciation training by using speech recognition to evaluate learners and assist them in correcting mistakes [11]. Other systems have used virtual simulations of real-life contexts to facilitate learning [20]. We demonstrate that our mnemonic keyword generation system can assist vocabulary learning and can thus be integrated with other CALL methods in comprehensive language learning systems.

LINGUISTICS TERMINOLOGY

Although TransPhoner keywords can be used for a variety of tasks, for clarity of exposition we focus on foreign language learning and use conventional terminology from linguistics. A *foreign* language is one in which a learner is not completely fluent. *Native languages* are the languages in which a learner has attained fluency, usually at an early age. Each language has at least one associated writing system, which we call a *script*, and a spoken language which obeys a set of phonological rules.

We represent pronunciations using the International Phonetic Alphabet (IPA) notation [19]. The IPA organizes and categorizes a comprehensive inventory of vowels and consonants exhibited in human speech and can be used as a language-agnostic transcription of *phones*: the basic sound units of speech. Each language uses a different subset of phones conceptualized as language-specific *phonemes*: the basic units of the phonology of that language that can distinguish meaning.

Phonemes are combined to form pronunciations of words. A phoneme can be viewed as a set of phones considered equivalent under the phonetics of a particular language. For example, the English phoneme /k/ occurs in words such as "kit" (IPA: [k^{h}ɪt]) and "skill" (IPA: [skɪl]) but the physical sound (i.e. phone) is different. "Kit" is pronounced with the phone [k^{h}] (aspirated k) whereas "skill" is pronounced with [k] (unaspirated k). In English these are so called *allophones* but in other languages such as Thai, they represent different phonemes and can be used for semantic distinctions.

Analogously, a *grapheme* is the smallest semantically distinguishing unit in a written language. Examples of graphemes include alphabetic letters in alphabets (e.g. a, ω), ideograms such as Chinese characters (e.g. 木, 字), and syllabic characters such as the Japanese kana (e.g. あ, ヒ).

DESIGN PRINCIPLES

We look at prior research on effective mnemonic keywords to lay out a set of design principles for a keyword generation system. At a high level, the mnemonic keywords should be memorable and have high reminding power for both the foreign word and the native translation. This requirement implies the following desirable properties for the generated keywords:

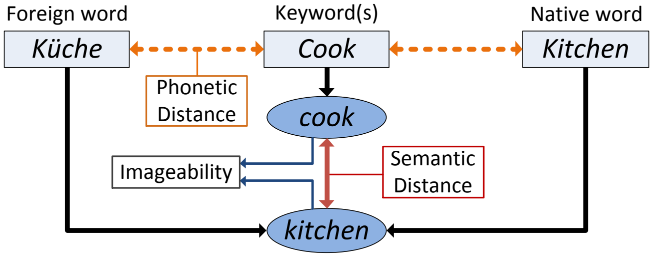


Figure 2. Illustration of the mnemonic keyword learning strategy. Effective keywords are phonetically similar to the foreign word, correspond to highly imageable concepts, and are semantically close to the concept. Concepts in dark blue circles, word forms in light blue boxes.

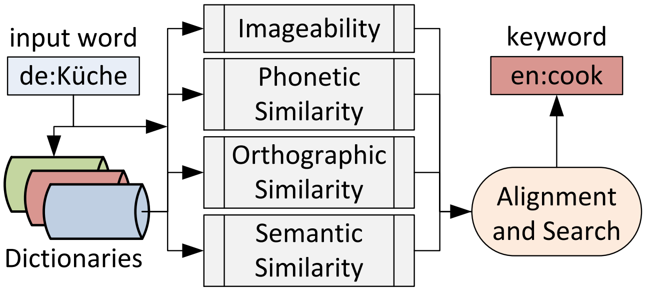


Figure 3. The TransPhoner system architecture: input words in a given foreign language are looked up and evaluated against words in a target native language using several similarity and quality measures. An alignment and search algorithm optimizes suggested keyword choice to give highly imageable, phonetically and semantically proximal keywords to be used as mnemonics for the input word.

Language	Dictionary	IPA Pronunciations	# Words
English	Scrabble list[2]	`www.dictionary.com`	63,261
French	FreeDict[3]	FreeDict	7,452
German	FreeDict	Rule-based[4]	80,235
Japanese	EDICT[5]	Rule-based	213,829
Mandarin	CEDICT[6]	CJKlib[7]	104,528

Table 1. Dictionary data used by our system. Dictionaries provide word lists and definitions. IPA pronunciations are retrieved from pronunciation dictionaries, or by using rule-based systems for languages with regular grapheme-to-phoneme mappings.

- *Phonetically similar*: aid recognition and generation of pronunciation
- *Highly imageable*: improve memorability and learnability
- *Semantically related*: help associate word form to concept

Our goal is to demonstrate that we can computationally generate keywords that are effective for learning vocabulary and guiding pronunciation. There are several dimensions along which a system such as this can be adjusted to generate better results for particular tasks. However, the focus of our paper is to present results utilizing simple, sensible defaults.

The architecture of our system is illustrated in Fig 3. Information for input words is retrieved by searching available dictionaries. Measures of word imageability, phonetic similarity, orthographic similarity and semantic similarity are then used to evaluate candidate keywords from the appropriate target language dictionary. The results are combined in a joint objective function and a search algorithm is used to optimize for the best candidate keyword.

LANGUAGE RESOURCES

Before applying a computational method to our problem, we need data sources for the orthography, definitions and pronunciations of words in all desired source and target languages. We retrieve this data from dictionaries (see Table 1). We use the IPA as a common representation for pronunciations to facilitate cross-language comparison. The IPA transcriptions were annotated with syllable separators (syllabification) using a rule-based system for English[1] and French, or induced directly during the rule-based transcription to IPA for German, Japanese and Mandarin.

We also need to reason about word semantics, and to connect word concepts between languages. For the former, we use WordNet [29], a lexical ontology with senses (meanings) for English words. Many words have multiple senses—an example is "bank" which can refer to a financial institution or to a river bank. For the latter, we use the Universal WordNet (UWN) [10], which connects other languages to the English WordNet. UWN also provides a weighted score for the degree to which a given word reflects each candidate sense. From the senses of each word, we retrieve definitions for computing imageability scores and semantic similarity between words.

SIMILARITY MEASURES

Orthographic similarity is useful for languages with similar scripts, since associating the written word forms in two languages can be effective for learning. However, phonetic similarity is more important in general as it helps in remembering the pronunciation of a word, and in avoiding pitfalls due to differences of script-to-phonetic mappings between languages. For example, "ch" is pronounced /ç/ in German, similar to the initial consonant in "hut", rather than English "chat".

An imageability measure is critical for selecting memorable keywords that are easy to associate with the word concept being learned. Finally, semantic similarity to the target word concept is useful for facilitating the forming of associations between foreign words and their native language translations.

Orthographic Similarity

A simple way to measure similarity of written form is using orthographic distance. For words in the same script we use a simple edit distance (Levenshtein distance) between their graphemes (letters). The Levenshtein distance [26] counts the number of operations (insertions, deletions and substitutions) that are required to transform one string into the other.

Phonetic Similarity

To compute phonetic similarities between pronunciations of words in different languages we consider each word as a sequence of phones from that word's IPA transcription. Again, we use a Levenshtein distance, now with a weight matrix defining the cost of substituting one phone for another, to compute the phonetic distance between two words. We search for an alignment (mapping of phones in one word to those of another) such that the Levenshtein distance is minimized.

Many phonetic similarity methods exist—a survey is provided by Kondrak [24]. We use the phonetic similarity defined by the ALINE phonetic alignment system which has been shown to align cognates between languages [23]. This scheme categorizes vowels and consonants in separate feature spaces where each phone is represented as a vector of values. Each feature dimension has an associated weight used in computing the overall distance between phones.

[1] `github.com/codyrobbins/syllabify`

[2] `code.google.com/p/scrabblehelper/`

[3] `www.freedict.org/en/`

[4] `korpling.german.hu-berlin.de/~amir/Text2IPA_German.pl`

[5] `www.csse.monash.edu.au/~jwb/edict.html`

[6] `cc-cedict.org/wiki/`

[7] `code.google.com/p/cjklib/`

We simplify the ALINE algorithm by omitting phone expansions (i.e. no special costs for matching two phones to one, or one phone to two). We handle matching of syllable breaks by adding a constant match score $C_{sep} = 20$, and a mismatch cost $C_{skip} = -10$, as defined by ALINE. The phone feature set used by ALINE is not complete—a few phones are considered equivalent (e.g. /i/, /ɪ/), so we add an extra weight for preferring exact phone matches[8].

Word Imageability

Unlike the other similarity components, methods for computing word imageability are largely unstudied. We describe a simple approach which aims to take into account word familiarity and easiness of acquisition as proxies for word imageability. Though not our focus here, an investigation and evaluation of approaches for computing imageability is an interesting avenue for future work.

A large corpus of imageability ratings for words is not available. However, studies in cognitive psychology have shown that word imageability is highly correlated with age of acquisition (AoA), an estimate of the average age at which children acquire a word [5, 16, 28, 33]. We therefore use a corpus of AoA ratings for more than 50,000 English words by Kuperman et al. [25], as well as other word features such as familiarity and part of speech to compute imageability ratings. We then propagate imageability values to other languages using the inter-language mappings from UWN.

We begin by estimating the imageability of English words in the Kuperman corpus of AoA values using a linear regression method. Additional word features for this regression come from the MRC Psycholinguistic Database [37] which contains linguistic and psycho-linguistic attributes of English words. Importantly, it contains imageability values for 4579 words derived from several previous corpora [7, 8, 16]. We normalize this imageability to be between 0 and 1, and then train a linear regression model using AoA, familiarity, and part of speech feature values from the Kuperman corpus.

We then compute the imageability of foreign words and English words not covered by the Kuperman corpus from words with known imageability using a two-step averaging scheme:

$$I_W(w) = \sum_{(a_i,s_i)\in \mathrm{S}(w)} \frac{a_i I_S(s_i)}{\sum_j a_j} \quad ; \qquad I_S(s) = \sum_{(b_i,w_i)\in \mathrm{W}(s)} \frac{b_i I_W(w_i)}{\sum_j b_j}$$

where $I_W(w)$ and $I_S(s)$ are the imageability values of a word w and sense s respectively, and $a_i; b_i$ are the word-sense association weights from UWN. Words not covered by UWN are given an imageability value of 0, under the assumption that rare words will be unfamiliar to most people, and consequently are unlikely to be highly imageable. Figure 4 plots a histogram of the resulting imageability values.

Semantic Similarity

We use a simple bag-of-words [27] (BoW) model over WordNet sense definitions to approximate relatedness of word concepts. We represent each sense as a vector of counts of words occurring in the sense definition. We filter the BoW vector to include only nouns, verbs, adjectives and adverbs. Part of speech tagging is done using the Stanford POS tagger [35]. The BoW vector for a word is then represented as a weighted sum of the BoW vectors for each sense of that word. The similarity between two words is computed using the cosine distance between the BoW representations: $sim(w_1, w_2) = v_{w_1} \cdot v_{w_2}/\|v_{w_1}\|\|v_{w_2}\|$ where $v_w = \sum_{(a_i,s_i)\in \mathrm{S}(w)} a_i v_{s_i}$ and v_{s_i} is the BoW vector representation of sense s_i.

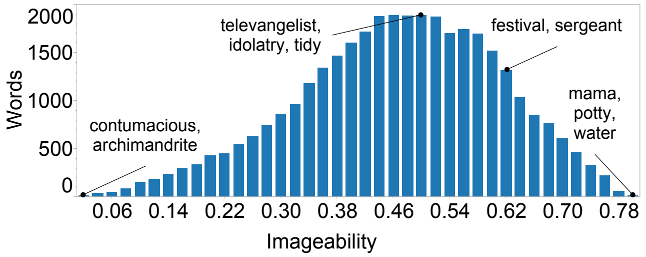


Figure 4. Histogram of imageability for 35,000 English words. Example words annotated. Imageability ranges between 0 and 1 on the horizontal.

We chose to use a simple approach for approximating semantic similarity—many alternatives exist from extensive work in Natural Language Processing [36]. Similarity measures over the WordNet hierarchy, or more advanced methods such as Latent Semantic Analysis or word embeddings trained with neural networks are easy to incorporate.

FINDING CANDIDATE KEYWORDS

We combine the similarity measures in an objective function and search for good keywords given particular input words.

Combined Objective Function

Since we do not constrain our system to single words on either input or output side, we consider sequences of words for our objective function. Given a sequence of words in a foreign language W_f as input, we find the sequence of words W^* from all candidate target native language word sequences W_n such that it maximizes a weighted combination of phonetic similarity σ_p, imageability I, semantic similarity σ_s, and orthographic similarity σ_o.

$$\begin{aligned} W^* = \underset{W_n = w_1,\ldots w_n}{\operatorname{argmax}} \quad & \alpha_p \sigma_p(W_n, W_f) + \alpha_o \sigma_o(W_n, W_f) \\ + \quad & \alpha_s \sigma_s(W_n, W_f) + \alpha_i I(W_n) \text{ where} \end{aligned}$$

$$\begin{aligned} \sigma_p(W_n, W_f) &= -\mathtt{LevDist}(\mathtt{phones}(W_n), \mathtt{phones}(W_f)) \\ \sigma_o(W_n, W_f) &= -\mathtt{LevDist}(\mathtt{chars}(W_n), \mathtt{chars}(W_f)) \\ \sigma_s(W_n, W_f) &= \mathtt{sim}(\mathtt{BoW}(W_n), \mathtt{BoW}(W_f)) \end{aligned}$$

$$\mathtt{BoW}(W) = \sum_{i=1}^{n} \mathtt{BoW}(w_i); \quad I(W) = \sum_{i=1}^{n} I(w_i)$$

The components of the objective function each assign a quality score to a candidate keyword. The relative importance of these components and the independent efficacy of each at aiding language learning is an interesting research question

[8] ALINE output values are divided by 100 to make them comparable with similarities from other components.

Input	English	German	French	Japanese	Mandarin
en:watermelon 🔊 /ˈwɔ.təɹ.ˌmɛ.lən/ watermelon	water million 🔊 /ˈwɔ.təɹ ˈmɪ.lyən/ water, million	Rotte mahlen 🔊 /ʀɔ.tə maː.lən/ herd, grind	voiture main lune 🔊 /vwa.ˈtyʁ mɛ lyn/ car, hand, moon	夜手メゾン 🔊 /jo te me.zoɴ/ night, hand, house	苜縢密林 🔊 /mu˥˩ təŋ˧˥ mi˥˩.lin˧˥/ clover, flying dragon, jungle
de:Gesundheit 🔊 /gə.zʊnt.haɪt/ health	gazette height 🔊 /gə.ˈzɛt haɪt/ gazette, height	Gesamtheit 🔊 /gə.zamt.haɪt/ entirety	goût zone chatte 🔊 /gu zon ʃat/ relish, area, cat	偽善者手 🔊 /gi.zeɴ.ʃa te/ hypocrite, hand	革新海 🔊 /kɤ˧˥.ɕin˥ xai˨˩˦/ innovation, sea
ja: お早う 🔊 /o.ha.joː/ Good morning	Ohio 🔊 /oʊ.ˈhaɪoʊ/ Ohio (state)	Ohrhörer 🔊 /oːɐ.høː.ʀɐ/ earphones	eau ration houx 🔊 /o ʁa.ˈdjo u/ water, ration, holly	オヒョウ 🔊 /o.hjoː/ halibut (fish)	五花肉 🔊 /u˨˩˦.xua˥.ʐou˥˩/ pork belly (food)

Table 2. Representative TransPhoner keyword results between several pairs of languages. Input word on the left, with source language indicated by the prefix. When targeting the same language, the input words are removed from the output candidate list. See supplemental materials for more results.

which is beyond the scope of this paper. For the examples presented here, we found $\alpha_p = \alpha_s = \alpha_i = 1$ and $\alpha_o = 0.5$ to provide good results.

Search Algorithm

To efficiently search over candidate matches, we use dynamic programming with either a phone-based trie or a character-based trie [14]. Tries are useful because we search over word sequences. For single word outputs, iterating over all words in a dictionary would suffice.

We take all words in the target language and create a phone-based trie. Potential phone matches are taken from the position in the trie. The cost of the match is the phonetic distance between the phones. This allows for re-using the computed Levenshtein distance between words with common prefixes. Whenever a word is matched, we loop back to the root of the trie, adding a syllable break between words. We also add the cost for selecting the matched word, incorporating non-phonetic similarity costs.

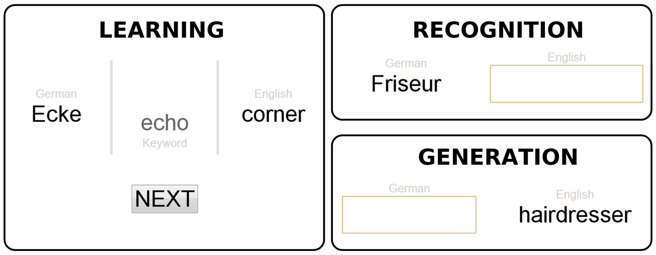
At each stage, we use a beam of N-best choices to ensure that we do not prune choices which may not be the closest phonetic match but have high scoring imageability or semantic similarity to the foreign word. This also allows us to generate N-best keyword lists from which a user can choose.

EXAMPLE OUTPUT KEYWORDS

Table 2 gives some example results of top keyword sequences suggested by TransPhoner when given the input words in the leftmost column. Output keywords are given in all five languages for which we have dictionary data. Keyword output is achieved at near-interactive rates, usually on the order of a second per source-target language pair, largely determined by the size of the target language dictionary.

EVALUATION EXPERIMENT

To evaluate the quality of keywords generated by TransPhoner we compared against manually selected keywords in the concrete application context of foreign language vocabulary learning. Prior work by Ellis et al. [13] has shown that manually chosen keywords can facilitate vocabulary learning. Our experimental goal was to show that we can generate equally effective keywords in this scenario with a direct evaluation metric: recall of learned words.


LEARNING
German
Ecke
echo
Keyword
English
corner
NEXT

RECOGNITION
German
Friseur
English

GENERATION
German
English
hairdresser

Figure 5. Screenshots of our vocabulary learning study interface.

Vocabulary Learning Study

We hypothesized that presenting TransPhoner keywords to people while they learn vocabulary improves learning performance by increasing new word retention. Ideally, this improvement would match or exceed the one imparted by manually chosen mnemonic keywords. To test this hypothesis, we carried out a vocabulary learning study. We used the evaluation set of 36 German words and manually chosen keywords by Ellis et al. as a comparison baseline. In addition, we randomly sampled from the 25000 most frequent English words to create a *random* keyword control condition. We thus have four conditions for keywords: *none*, *random*, *manual* and *TransPhoner*. As an example, for the German word “Friseur” (hairdresser), the *random* keyword was “opal”, the manually chosen keyword by Ellis et al. was “freezer”, and the automatically generated TransPhoner keyword was “frizzy”.

Participants

We recruited participants from the Amazon Mechanical Turk workplace. Participants were required to not have any exposure to German and to be fluent in English. In total, we recruited 100 participants (60 male) with an average age of 32.3 years ($SD = 9.9$). Participants were compensated with $1.70 for the study and told a $0.10 bonus is available for average scores higher than 70%. To account for workers who did not earnestly attempt the task, we filtered any participants with average scores lower than 25%. This leaves 19 participants each for *none* and *manual* conditions, and 18 each for *random* and *TransPhoner* conditions.

Procedure

Fig 5 shows our web-based study interface. We designed the interface to match the experimental procedure of Ellis. Participants were randomly assigned to one of the four conditions.

They first took a short demographic survey and read instructions introducing the task. In keyword conditions, participants were instructed to do their best to *"imagine a visual scene connecting the given keyword with the English meaning, and the sound of the German word"*.

The study proceeded in three phases: learning, recognition and generation. First, participants were shown a block of 12 word pairs and asked to memorize the association. Keywords were shown in the center (except in the *none* condition). Words were pronounced twice, two and seven seconds after being shown. When the participant was ready, they proceeded by clicking a next button. The screen was blanked for one second, and the next word was shown. Order of presentation was randomized between participants.

Next, participants had to input the English meaning of the same 12 German words (re-randomized order) in a recognition phase. The word was again pronounced twice at the same intervals. Participants had a total of 10 seconds for input, after which the screen was blanked for one second and the next word was shown. Finally, participants were asked to give the German word for each English translation in the generation phase, spelling the German words as best as they could.

After one stage of learning–recognition–generation, the procedure was repeated with two more blocks of 12 words for a total of 36 words. Participants were instructed that the first stage was for training while the other two would be scored.

Design

The experiment was a mixed between- and within-subjects factorial design with the keyword condition {*none*, *random*, *manual*, *TransPhoner*}, participant {1...74}, word {1...36} and task {*recognition*, *recall*} as factors. All participants provided a recognition and generation response for each of 36 words for a total of 5328 responses.

The dependent measure was the recall score computed by comparing participant responses with the correct translation. We allowed for imperfect spellings by using the Levenshtein distance between participant response and the correct word, dividing by maximum possible distance and subtracting from 1 to create a normalized score between 0 and 1. Scores above 0.5—corresponding to at least half of the characters matching—were given partial credit equal to the score, while lower values were considered incorrect. Common synonyms such as "pants" and "trousers" were also considered correct. At the conclusion of the task, we also asked participants in the keyword conditions to judge the keywords for helpfulness in learning foreign vocabulary on a 5-point Likert scale, as well as to provide optional comments on the task.

Study Results

Vocabulary learning scores

Table 3 summarizes the average participant scores and keyword helpfulness ratings across conditions. The *none* condition had the lowest scores, followed by *manual*, *random* and finally the *TransPhoner* condition with a combined score of 76.1%. Interestingly, the difference between *random* and *manual* conditions is small, an observation we discuss later, in light of participant comments.

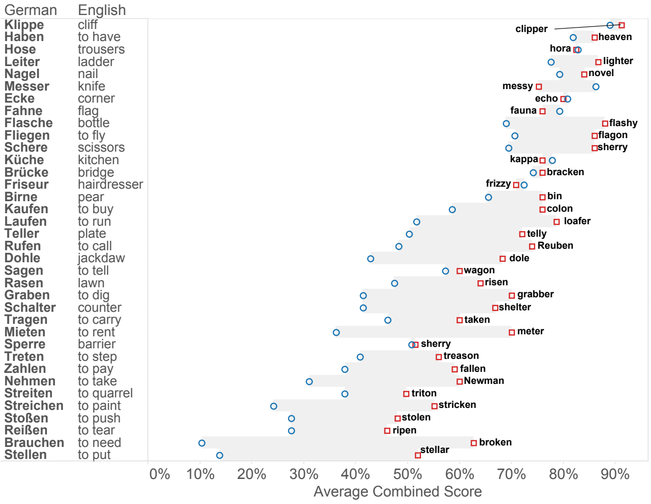


Figure 6. Comparison of average participant scores in conditions *none* (blue circles) and *TransPhoner* (red squares with keyword) for all words. Words ordered by decreasing mean score (easiest to hardest).

Fig 6 plots participant score improvement per word between the *none* and *TransPhoner* conditions. TransPhoner keywords significantly improve learning, with the effect being more pronounced for harder words. In general, verbs are harder to learn, a result agreeing with prior work [13].

Keyword helpfulness ratings

The mean keyword helpfulness ratings (see Table 3) were lowest for *random*, followed by *TransPhoner* and *manual*. All pairwise keyword rating differences were significant under Wilcoxon rank-sum tests with Bonferroni-Holm correction ($p < 0.001$ for all, except *manual–Transphoner* $p < 0.05$). Random keywords were generally disliked—an observation reflected in comments by study participants. The better performing *manual* and *TransPhoner* conditions had polarized ratings (almost purely 1 or 5 ratings). This polarization is not unexpected since some participants can have low affinity for keyword-based vocabulary learning, an observation also reflected in participant comments.

Keyword similarity scores

Table 3 also reports the normalized similarity scores between keyword and target word for all 36 German words used in the study. Unsurprisingly, randomly selected keywords had the lowest score along all dimensions. TransPhoner keywords had the highest overall (0.87) and phonetic (0.84) similarities, while manual keywords had higher semantic (0.25) similarity and imageability (0.65). Overall similarity and phonetic similarity were found to have a statistically significant positive correlation with combined learner scores (Pearson's $r(3526) = 0.06, p < 0.01$ and $r(3526) = 0.8, p < 0.001$ respectively). We did not find a significant correlation between either semantic similarity or imageability and the learner scores.

Further investigation of the effect of different forms of similarity between keyword and target word on learner perfor-

Condition	Rec %	Gen %	Comb %	Helpfulness	Overall Sim.	Phonetic Sim.	Semantic Sim.	Imageability
None	60.6	60.7	60.7	—	—	—	—	—
Random	68.1	66.9	67.5	2.09	0.40	0.38	0.11	0.32
Manual	69.3	64.4	66.8	**3.78**	0.78	0.72	**0.25**	**0.65**
TransPhoner	**76.4**	**75.9**	**76.1**	3.46	**0.87**	**0.84**	0.15	0.51

Table 3. Left side: average participant recognition, generation and combined vocabulary learning scores, and keyword ratings. Right side: average normalized target word to keyword similarity and imageability scores.

mance is an interesting avenue for future work. We noted that with English as a target language, the phonetic similarity dimension tends to strongly constrain word choice. In contrast, in target languages with higher phonetic-to-semantic multiplicity, such as Mandarin and Japanese, jointly optimizing phonetic similarity along with semantic similarity and imageability becomes easier.

Statistical analysis

We standardized all continuous numeric data by subtracting the mean and dividing by the standard deviation. Standard ANOVA does not account for per-word and per-participant variation leading to increased risk of type II errors. We therefore used mixed effects models which support both fixed effects and random effects (such as stimulus word and participant), and which are commonly used in psycholinguistics [3]. A good introduction for the HCI community is provided by recent work in machine translation post-editing [17]. Following this work, we also report significance results using likelihood-ratio (LR) tests—a measure equal to twice the difference between the log-likelihood of the model and the null hypothesis.

We fit a mixed effects model with keyword condition and logarithm of time spent learning the word as fixed effects, and both word and participant as random effects[9]. There was a significant main effect of the keyword condition on average participant score $\chi^2(3, N = 5184) = 7.87, p < 0.05$. We performed follow-up pairwise Welch's t-tests with Bonferroni-Holm correction between all keyword conditions. The mean score differences were all significant at $p < 0.001$, except *manual–TransPhoner* at $p < 0.01$, and *random–manual* (not significant $p = 0.07$). All keyword conditions performed better than no keywords, including *random* likely due to participants being primed to use a mnemonic strategy.

The absence of a significant difference between manual and random keywords seems surprising. However, we note that we used the *manual* condition keywords in a different way than Ellis et al., since the original work prompted participants with complete sentences containing the keywords (which were emphasized). Furthermore, the *random* condition likely resulted in participants reverting to a strategy of coming up with their own keyword. This hypothesis is supported by comments from *random* condition participants who complained about the quality of the keywords and stated that they came up with their own keywords. Prior work has shown that even when not asked to use keywords, participants come up with their own and use them effectively, which makes it difficult to control the participant learning strategy [13].

Participant comments

Though comments at the conclusion of the study were optional, more than half of our participants provided them. Comments from *manual* and *TransPhoner* condition participants were overwhelmingly positive indicating that they thoroughly enjoyed the task, and would like to continue taking similar experiments. Some examples include: *"Very interesting way for learning a new language"*, *"keywords really helped"*, *"I have always wanted to learn German but this HIT has really opened up that it might be really hard but it could be done by me"*. One of the negative comments mentioned the potential interference effect of keywords with target word meaning: *"I remembered them far better than I remembered the meaning of the word for some reason"*.

Despite improving performance over no keywords, the *random* keyword condition garnered largely negative comments: *"I didn't think the keyword helped in most cases – better off trying to make the sound fit with the english word"*, *"I actually felt like the keywords threw me off a bit"*, *"I do better when I make my own connections."* and *"The keywords made it harder for me. I like to make up my own keywords. For someone who didn't do that already, it might help"*. The latter comments indicate that some participants in the *random* condition compensated by creating their own keywords. A similar comment even occurred in the *none* condition: *"This was very difficult even using word association"*, reflecting that participants may use keywords even when not prompted to do so.

Summary

Our results are consistent with prior studies in keyword-based vocabulary learning. Short-term recognition and generation are improved by TransPhoner keywords, with the effect being larger for recognition. Though detailed evaluation of keyword-based learning is beyond our scope, and covered by much prior work, these results indicate that automatically generated keywords can significantly facilitate vocabulary learning, with performance matching or exceeding manual keywords. We note that our experiment only tested short-term effects. However, a retention study over a period of 10 years has shown that the positive effect of keywords carries over into long-term retention [4].

OTHER APPLICATIONS

So far, we focused on generating mnemonic keywords for foreign vocabulary learning. However, the TransPhoner system can be used in various other scenarios as well. Examples include: retrieval of images for generated keywords, transliteration within one language for clarifying pronunciation or

[9] Other independent variables such as age and gender had no significant effect, and interactions between keyword condition and learning time were not found. We used the `lme4` R package and optimized fit with maximum log-likelihood [3].

SAT Word	Meaning	Keywords
camaraderie	mutual trust and friendship	camera diary
subservient	obey others unquestioningly	sub servant
querulous	complaining in a whining manner	Queen rule loss
nonchalant	feeling or appearing casually calm	noon shall law neat
vicissitude	change of circumstances or fortune	visit side dude

Table 4. SAT vocabulary and example mnemonic phrases from top 10 results generated by TransPhoner.

Input	"Slanguage" Books	TransPhoner
fr:Merci beaucoup 🔊 /mɛʁ.ˈsi bo.ˈku/ Thank you very much	Mare-see-bow-coo 🔊 /mɛər si boʊ ˈkɔː/ $\sigma_p = 0.82, \sigma = 0.92$	Meg-see-boo-coup 🔊 /mɛg si bu ku/ $\sigma_p = 0.83, \sigma = 0.90$
ja: 楽しむ 🔊 /ta.no.ʃi.mu/ To enjoy	Ton-know-she-Moo 🔊 /tʌn noʊ ʃi mu/ $\sigma_p = 0.80, \sigma = 0.89$	Ta-gnaw-she-moo 🔊 /tɑ nɔ ʃi mu/ $\sigma_p = 0.90, \sigma = 0.93$
zh: 薄烤饼 🔊 /po˧˥.kau˨˩˦.piŋ˨˩˦/ Pancake	Bow-cow-bing 🔊 /boʊ kaʊ bɪŋ/ $\sigma_p = 0.51, \sigma = 0.65$	Paw-cow-ping 🔊 /pɔ kaʊ pɪŋ/ $\sigma_p = 0.80, \sigma = 0.87$

Table 5. Some phrases with transliterations provided by "Slanguage" guide books [12], and corresponding transliterations by TransPhoner. Normalized phonetic (σ_p) and overall (σ) similarity values between input and output given for comparison.

Original	Food Topic	Gambling Topic
Sweet dreams are made of this	Sweet creams are made of this	Sweet deals are made of this
Who am I to disagree	Who am I to daiquiri	Who am I to lottery
I travel the world and the seven seas	I travel the world and the saffron seas	I raffle the world and the seven seas
Everybody's looking for something	Everybody's looking for dumpling	Everybody's looking for gambling

Table 6. Example of phonetic transformation constrained by topic: lyrics from the song "Sweet Dreams" by Eurythmics are transformed to be semantically closer to the topics of "food" and "gambling".

creating mnemonics for complex words, and pure phonetic transliteration of tourist guide book phrases.

Keyword Images

Providing images for the generated keywords may facilitate the formation of a mental image. Though we do not empirically evaluate this hypothesis, we can easily retrieve relevant images for keywords from the web. TransPhoner currently uses the Google Image API to demonstrate this functionality (examples shown in Fig 1).

Complex Word Mnemonics

By matching rare complex words against shorter, more frequent keywords in the same language, we can create mnemonic keyword phrases for learning the complicated words. This is similar in spirit to the re-spelling work of Hauer et al. [18], using complete words instead of syllables, and optimizing for memorability as well as phonetic similarity. Table 4 gives examples for words taken from a list of vocabulary covered by the Scholastic Aptitude Test (SAT) exam, commonly taken by high school students in the United States.

Pure Phonetic Transliteration

When learning foreign word pronunciations, transliterations formed with the phonetically closest keywords in the target language can be effective as pronunciation guides. In fact, such keyword transliterations are designed manually and marketed as pronunciation learning guide books [12]. Table 5 shows some example phrases found in this series of books, the transliterations provided by the authors, and corresponding transliterations generated by TransPhoner. In general, TransPhoner transliterations are at least as close to the input language pronunciation as the manually designed phrases. Most importantly, TransPhoner generates these transliteration suggestions in seconds, whereas the effort in authoring them from scratch is considerable.

Topic-based Homophonic Transformation

TransPhoner can be used to automatically suggest topic-based homophonic transformations of word sequences. An application of this is in creating novel, phonetically similar interpretations of song lyrics–usually for comedic effect–known as *soramimi* or *mondegreen*. An example is in Table 6. We use all WordNet hyponyms under a given topic as candidate replacement words and for each phrase we compute phonetic similarity between original words and all candidates, offering the top 5 closest matches as suggestions.

LIMITATIONS AND FUTURE WORK

In our implementation we relied on dictionary data sources for phonetic information. A more robust approach could handle out-of-dictionary words by estimating pronunciation through grapheme to phoneme conversion systems [6]. Such systems would also help us deal with word pronunciations which are context-sensitive. For example, the Japanese "人" for person can be pronounced as /hi.to/, /d͡ʑiɴ/ or /niɴ/ depending on the context of the character. We have also not investigated how to transfer the prosody and pitch contours of utterances from one language to another. These attributes are particularly important for tonal languages such as Mandarin.

Another avenue for future work is generation of longer phrases and incorporation of language models to form grammatically valid sentences. An interesting empirical question is whether keyword sequences which are also valid sentences are better for learning. Richer stimuli such as context sentences connecting keyword and foreign word, related images, and animations could be integrated with such a system to further facilitate learning.

Finally, methods for tuning the generation of keywords for different tasks and types of users would be valuable to explore. For instance, we may tune keywords for younger learners by restricting them to more basic and imageable words, through an age of acquisition threshold.

DISCUSSION

Benefit of mnemonic keywords

We empirically saw that TransPhoner keywords improve short-term vocabulary learning, a result in agreement with existing research. Why are keywords helpful for learning? One hypothesis advocated by prior work is that keywords make learning more fun and engaging. For instance, the initial phases of foreign language learning require absorbing thousands of new words and can be discouraging, especially when approached mainly through rote repetition. Likewise, memorization of complex terminology for academic topics is intimidating to many students.

Though rote memorization is effective for learning, it is unfortunately also monotonous and discouraging, especially given the unfamiliar phonetics and orthography of a new language. Keywords and visual imagery can make the process more enjoyable, especially for younger learners. As such, keywords are a tool to surmount the initial learning hump and accelerate learning by facilitating associations at a point when the learner lacks knowledge necessary to form them independently. Keywords provide a learning scaffold that can engage learners and can also inspire them to come up with their own mnemonics.

The comments from our study indicate that participants enjoyed learning when keywords were carefully selected, and want to continue learning with keywords. Several participants explicitly requested to be informed when more similar learning tasks are available. In contrast, participants disliked random keywords and indicated their unhappiness with the task. Since the difference between these two conditions was the choice of keywords, we see that well chosen keywords can strongly engage and motivate learners, and that badly chosen keywords can have the opposite effect.

Implications for practical learning

Much research has shown that keywords are effective learning aids. However, application to actual learners and classrooms has been restricted. Teaching materials with keywords are scarce—the LinkWords system being a notable exception.[10] Creating such material requires considerable time and effort. Furthermore, each student responds differently to particular keywords, so success is largely contingent on tailoring them to the student's learning affinities. Therefore, a system for suggesting candidate mnemonic keywords can be invaluable for making keyword learning practically feasible and broadly accessible. Used in conjunction with other learning strategies, such a system can have tremendous positive impact.

CONCLUSION

We have presented TransPhoner, a cross-lingual system for generating mnemonic keywords. We evaluated TransPhoner keywords with a web-based vocabulary learning study and found that they lead to improved recall, with the effect being more prominent for harder words. Investigating the efficacy of TransPhoner keywords for specific learner demographics, and measuring the degree to which recall improvements lead to longer term retention are both interesting avenues for further research.

We also demonstrated how TransPhoner keywords are useful in a variety of other scenarios: as mnemonics for aiding learning complex vocabulary, as pronunciation guides for tourists and traveling professionals, and as suggestions for inspiring creative word plays.

The potential impact of mnemonic keyword systems is broad. In modern societies, knowledge of multiple languages is critical, so any improvement in learning languages can have significant long-range effects. Furthermore, the keyword mnemonic can easily be applied to other kinds of information such as country capitals or elements of the periodic table.

We hope our results will inspire further work on methods for mnemonic-based teaching and learning. Beyond increasing learner performance, integration of keyword-based techniques in learning tools may benefit teachers by allowing for easier design of effective teaching materials. Finally, psycholinguistics researchers studying mnemonic keyword learning strategies stand to benefit from a system that can generate keywords with different phonetic and semantic properties.

We have barely scratched the surface of the multi-faceted domain of mnemonic keywords. There is great potential for follow-up work on incorporating keywords in HCI systems to aid creative writing tasks, to provide on-the-spot pronunciation guides for tourists, and to assist people in learning a variety of material.

ACKNOWLEDGMENTS

We thank Gabor Angeli, Katherine Breeden, Matt Fisher and Daniel Ritchie for their help. The first author was supported by a Stanford Graduate Fellowship. The images in Figure 1 were released to the public domain by www.openclipart.com users: nicubunu, massimo, chrisdesign, palomaironique, chudq and stevelambert. The ratatouille image is by Flickr user Marcus Guimarães.

REFERENCES

1. Arnold, N., and Ducate, L. *Present and future promises of CALL: From theory and research to new directions in language teaching*. Computer Assisted Language Instruction Consortium, 2011.

2. Atkinson, R. C., and Raugh, M. R. An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory 1*, 2 (1975), 126.

3. Baayen, R., Davidson, D., and Bates, D. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* (2008).

4. Beaton, A., Gruneberg, M., and Ellis, N. Retention of foreign vocabulary learned using the keyword method: A ten-year follow-up. *Second Language Research 11*, 2 (1995), 112–120.

5. Bird, H., Franklin, S., and Howard, D. Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers 33*, 1 (2001), 73–79.

6. Chen, S. F. Conditional and joint models for grapheme-to-phoneme conversion. In *INTERSPEECH* (2003).

7. Clark, J. M., and Paivio, A. Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers 36*, 3 (2004), 371–383.

[10] www.linkwordlanguages.com

8. Cortese, M. J., and Fugett, A. Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers 36*, 3 (2004), 384–387.

9. de Groot, A. M., and Van Hell, J. G. The learning of foreign language vocabulary. *Handbook of bilingualism* (2005), 9.

10. De Melo, G., and Weikum, G. Towards a universal WordNet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, ACM (2009), 513–522.

11. Ehsani, F., and Knodt, E. Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm. *Language Learning & Technology 2*, 1 (1998), 45–60.

12. Ellis, M. *French Slanguage*. Gibbs Smith, 2012.

13. Ellis, N. C., and Beaton, A. Psycholinguistic determinants of foreign language vocabulary learning. *Language learning 43*, 4 (1993), 559–617.

14. Fredkin, E. Trie memory. *Communications of the ACM 3*, 9 (1960), 490–499.

15. Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., and Etkind, R. Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology 21*, 4 (2007), 499–526.

16. Gilhooly, K. J., and Logie, R. H. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation 12*, 4 (1980), 395–427.

17. Green, S., Heer, J., and Manning, C. D. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 439–448.

18. Hauer, B., and Kondrak, G. Automatic generation of English respellings. In *Proceedings of NAACL-HLT* (2013), 634–643.

19. International Phonetic Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

20. Johnson, W. L., Marsella, S., and Vilhjalmsson, H. The DARWARS tactical language training system. In *Proceedings of I/ITSEC* (2004).

21. Karimi, S., Scholer, F., and Turpin, A. Machine transliteration survey. *ACM Computing Surveys (CSUR) 43*, 3 (2011), 17.

22. Kern, R. Perspectives on technology in learning and teaching languages. *TESOL Quarterly 40*, 1 (2006), 183–210.

23. Kondrak, G. A new algorithm for the alignment of phonetic sequences. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, Association for Computational Linguistics (2000), 288–295.

24. Kondrak, G. Phonetic alignment and similarity. *Computers and the Humanities 37*, 3 (2003), 273–291.

25. Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods 44*, 4 (2012), 978–990.

26. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, vol. 10 (1966), 707.

27. Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to information retrieval*, vol. 1. Cambridge University Press Cambridge, 2008.

28. McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., and Lannon, R. An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental science 14*, 2 (2011), 181–189.

29. Miller, G. A. WordNet: a lexical database for English. *Communications of the ACM 38*, 11 (1995), 39–41.

30. Paivio, A. Mental imagery in associative learning and memory. *Psychological review 76*, 3 (1969), 241.

31. Paivio, A., Yuille, J. C., and Madigan, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology 76*, 1p2 (1968), 1.

32. Sagarra, N., and Alba, M. The key is in the keyword: L2 vocabulary learning methods with beginning learners of Spanish. *The Modern Language Journal 90*, 2 (2006), 228–243.

33. Stadthagen-Gonzalez, H., and Davis, C. J. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods 38*, 4 (2006), 598–605.

34. Takač, V. P. *Vocabulary learning strategies and foreign language acquisition*, vol. 27. Multilingual matters, 2008.

35. Toutanova, K., Klein, D., Manning, C., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. *NAACL* (2003).

36. Turney, P. D., Pantel, P., et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research 37*, 1 (2010), 141–188.

37. Wilson, M. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers 20*, 1 (1988), 6–10.