The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Learning to Find Translation of Grammar Patterns in Parallel Corpus

Kai-Wen Tuan[1], Yi-Jyun Chen[1], Yi-Chien Lin[1], Chun-Ho Kwok[1], Hai-Lun Tu[2], Jason S. Chang[1]
[1]Department of Computer Science, National Tsing Hua University
[2]Department of Library and Information Science,
Fu Jen Catholic University
{kevintuan,yijyun,nicalin,isaackch,helen.tu,jason}@nlplab.cc

Abstract

We introduce a method for assisting English as Second Language (ESL) learners by providing translations of *Collins COBUILD* grammar patterns (GP) for a given word. In our approach, bilingual parallel corpus is transformed into bilingual GP pairs aimed at providing native language support for learning word usage through GPs. The method involves automatically parsing sentences to extract GPs, automatically generating translation GP pairs from bilingual sentences, and automatically extracting common bilingual GPs. At run-time, the target word is used for lookup GPs and translations, and the retrieved common GPs and their example sentences are shown to the user. We present a prototype phrase search engine, *Linggle GPTrans*[1], that implements the methods to assist ESL learners. Preliminary evaluation on a set of more than 300 GP-translation pairs shows that the methods achieve 91% accuracy.

Keywords: Grammar Pattern, ESL Learning, Parallel Corpus

1 Introduction

In an era of globalization, English fluency becomes an increasingly important asset, and an increasing number of online services specifically target English as Second Language (ESL) learners. Dictionaries, thesauri, online English courses, and editorial tools are just a few examples. However, few if any of those services take into consideration the important relationship between grammar patterns (GPs) and word meanings. We expand upon the idea "one sense per collocation" proposed by Yarowsky (1993) and assume each GP would have only one word sense. For example, the word "keep" has multiple meanings: it meas "to delay or prevent" in the GP **V n from n** (e.g."keep candy from kids") and it means "to continue" in the GP **V -ing** (e.g."keep moving").

[1]https://linggle.com/

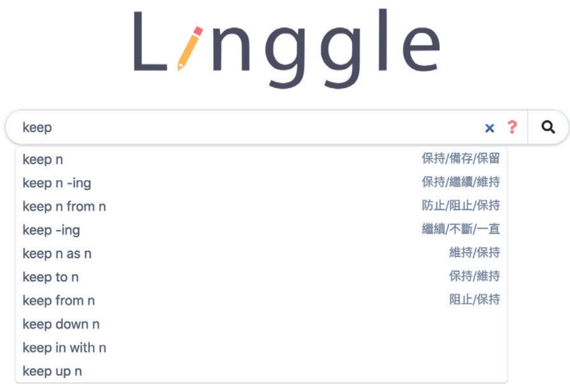

Figure 1: An example *GPTrans* search query *keep* : GPs, and Chinese translations

We focus our research on verbal GPs as linguistic researches identify verb phrases are particularly difficult for learners to understand. Moreover, verb phrase is a prominent component of sentence structure and lack of such knowledge often leads to grammatical errors.

We present a system, *Linggle Grammar Pattern Translator* (*Linggle GPTrans*), that automatically matches input words to corresponding GPs and relevent translations. Figure 1 shows the GPs and translations for the input word "keep." This system would help users to learn meanings of each word, in relation to its GPs.

2 Related Works

Language skills in English has proved indispensable along with the development of globalization. As a result, ESL learning has become an area of active research and many researches have worked on autonomous language learning (e.g., Kormos and Csizer (2014)).

Many researches show that, for non-native language learners, verbs are particularly difficult to learn compared to nouns (e.g., Hirsh-Pasek and Golinkoff (1999); Waxman and Booth (2001); Gleitman (1990); Gentner (1982); Imai et al. (2008)). In our system, we interactively provide the bilingual verb GP pairs to improve learning experience and efficiency.

In the past few decades, a large number of bilingual corpus resources have made statistical machine translation more and more feasible. In the 1990s, bilingual sentence alignment technology developed rapidly (Gale and Church (1991); Gale and Church (1993), Brown et al. (1991); Simard et al. (1993); Chen (1993)). Early research are aimed toward finding the corresponding bilingual sentences from bilingual corpus (Debili and Sammouda (1992); Kay and Roscheisen (1993)). Some studies use statistical models to improve the word correspondence generated by automatic alignment, such as Hidden Markov Model (HMM) (Brown et al. (1991)), log-likelihood ratio (Gale and Church (1991); Gale and Church (1993)) and K-Vec algorithm (Fung and Church (1994)). Based on the previous results, Melamed (1999) proposed the Smooth Injective Map Recognizer (SIMR), which regards bilingual phrase alignment as the best distribution of x-axis and y-axis in two-dimensional space. SIMR uses a greedy algorithm to calculate the best distribution of two-dimensional space as the calculation unit.

More recent researches concentrate on learning word translation and extracting bilingual word translation pairs from bilingual corpus, and then calculate the degree of mutual relationship between word pairs in parallel sentences, thereby deriving the precise translation (Catizone et al. (1989); Brown et al. (1990); Gale and Church (1991); Wu and Xia (1994); Fung (1995); Melamed (1995); Moore (2001)). Previously, we utilize statistical model in *Linggle* (Boisson et al., 2013), a linguistic search engine based on *Google Web 1T*. In our system, we focus on using statistical methods to extract translations of verbal GPs extracted from *Collin COBUILD Grammar Dictionary* (Cobuild et al., 2005).

(1) Parse sentences and extract grammar patterns
（Section 3.1）
(2) Extract translations of words
（Section 3.2）
(3) Count and filter headword translations for grammar patterns （Section 3.3）
(4) Extract Chinese pattern for grammar patterns
（Section 3.4）

Figure 2: Identification process

In the area of phrase alignment, Ko (2006) proposed a method for verb phrase translation. For specific verb fragments (e.g. *make a report to police*), automatic alignment is applied to calculate the collocation relationship across two language (e.g. when *make* and *report* appear together, *report* often corresponds to “報案” (bau an)), then word and phrase correspondences are generated (e.g. *make a report to police* correspond to “向警察報案”(shiang jing cha bau an)), to tally translations and counts. Chen et al. (2020) focus on the translation of noun+prepositional collocations. Using statistical methods to extract translations of nouns and prepositions from bilingual parallel corpora with sentence alignment, and then adjust the translations with additional information of Chinese collocations extracted from a Chinese corpus.

3 Methods

We attempt to identify the senses and translations of verbs in various GPs in *Collins COBUILD Dictionary*. Our identification process is shown in Figure 2.

3.1 Parsing Sentences and Extracting Grammar Patterns

In the first stage of the identification process (Step (1) in Figure 2), we parse each sentence and extract grammar patterns for each verb in each sentence. For example, the sentence ”*we will charge him with a crime.*” contains the verb *charge*, our goal is to extract the grammar pattern (GP) **V n with n** for the verb *charge* in the sentence, where **V** denotes the headword *charge*.

The input to this stage is English sentences from bilingual parallel corpora. We parse each

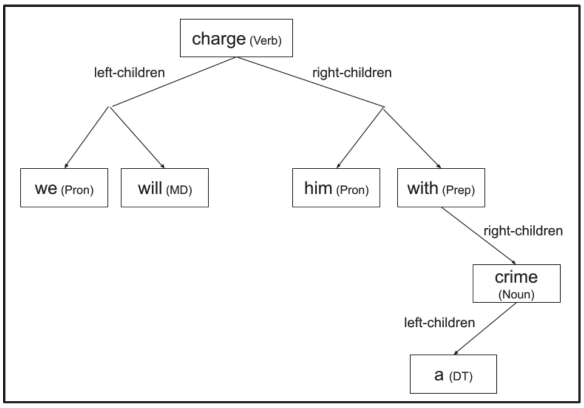

Figure 3: The structure of “we will charge him with a crime”

sentence into a tree structure to reveal the dependency of words in the sentence. For example, the sentence “*we will charge him with a crime*” will be parsed into a structure shown in Figure 3.

We use a recursive approach to extract GPs from the parse tree. We identify each verb, and consider all its right-children. We replace these words with their corresponding part of speech to form the GPs. Note that prepositions and specific function words (e.g. *out*, *up*, *down*) are not replaced. For example, in the sentence in Figure 3, the verb *charge* generate **charge n**. Then, the same process is then applied to the prepositions and other function words. The patterns generated from this process are then added to the partial GP to form a complete GP. For example, the prepositions *with* in Figure 3 generate a pattern **with n**. **with n** is then added to form a complete pattern, **charge n with n**. Note that there are some GPs with multiple word elements that can not be handle by the method described, such as **V wh-to-inf** and **V wh**. To handle these GPs, we design special rule to deal with these elements (e.g., **V wh-to-inf**).

Finally, we convert the generated GPs into the form consistent with *Collins COBUILD Dictionary*. In addition to replacing the headword in patterns as **V**, we also shorten the patterns that are too long to match GPs in *Collins Dictionary*.

The output of this stage is GPs which are in *Collins COBUILD Dictionary* of each verb in each English sentence.

3.2 Extracting Translations of Words

In the second stage of the identification process (Step(2) in Figure 2), we extract Chinese translations of each English verb. The input to this stage is English and Chinese sentences in bilingual parallel corpora with word alignment.

We use the method of extracting word translations in (Chen et al., 2020). For each English word, we consider its forward-correspondence to Chinese words and reverse-correspondence of these Chinese words, to filter out translations of the English word.

The output to this stage are Chinese translations for each English word. Sample translations of the word “charge” and the word ” keep” are shown in Table 1.

3.3 Counting and Filtering Translations

In the third stage of the identification process (Step (3) in Figure 2), we count and filter Chinese translation of headword for each grammar pattern(GP) and verb. For example, the verb *charge* has the GP **V n with n**. Our goal is to obtain the common translations such as ”控告” (kung kao, means ”accuse”) or ”指控”(jhih kong, also means ”accuse”) for *charge* in **V n with n**.

The input to this stage is English and Chinese sentences in bilingual parallel corpora with word alignment, GPs of each verb extracted from each English sentence in the first stage (Step(1) in Figure 2), and word translations extracted in the second stage (Step(2) in Figure 2).

For each GP of each verb, we compute the frequency of each translation as shown in Table 2. We than calculate the average and standard deviation in relation to other translation of the same GP. We filter and identify translations more frequent than average by 1 standard deviation.

To retain some less common translations which are correct (e.g., “跟上” (ken sheng, means “keep up with”) for V n of *keep*), we also compute the frequency of translations of verbs in each grammar pattern as shown in Table 5 and calculate the average and standard deviation. We filter and identify gram- mar patterns that are more frequent than average

Word	Translations
charge	主管 (1222), 負責 (599), 收費 (506), 押記 (358), 控罪 (324), 掌管 (280), 電荷 (266), 費用 (260), 收取 (258), 指控 (114), 充電 (66), 控告 (47), 罪名 (38), 檢控 (35), 徵收 (33), 控 (18), 指責 (11), 衝鋒 (11), 起訴 (6)
keep	保持 (1681), 維持 (312), 繼續 (304), 備存 (297), 不斷 (246), 保留 (236), 一直 (154), 儲存 (137), 保存 (61), 持續 (46), 保守 (42), 防止 (40), 不停 (39), 保管 (37), 留 (37), 保住 (36), 留住 (32), 阻止 (29), 存備 (12), 記錄 (11), 跟上 (11),

Table 1: Translations and count of “charge” and ” keep”

Word	GP*	Trans*	Count	Std	Example
keep	V n	保持	1669	**4.88**	keep space (保持距離)
keep	V n	維持	356	0.72	keep peace (維持和平)
keep	V n	備存	316	0.60	keep record (備存記錄)
keep	V n	繼續	147	0.06	keep a close watch (繼續密切留意)
keep	V n	記錄	52	-0.23	keep track of expenses (記錄支出)
keep	V n	飼養	33	-0.32	keep animal (飼養動物)
keep	V n	跟上	24	-0.33	keep pace (跟上步伐)

Table 2: Count the number of each translation for grammar pattern **V n** of “*keep*”.

by 1 standard deviation, and add the translation into the translation list of these grammar patterns for the verb.

The output of this stage is translations for each grammar pattern for each verb. For example, translations for each GP for the word *keep* and the word *charge* are shown in Table 4. After filter grammar patterns from each translation for each verb and add the translation into the translation list of these grammar patterns, we can extract translations that are relatively un- common such as “跟上”(ken sheng, means “keep up with”) or “飼養”(ssu yang, means “bread”) for “keep”.

3.4 Extracting Chinese patterns

In the forth and final stage of the identification process (Step(4) in Figure 1), we filter Chinese patterns for each English of each verb. For example, the verb ”*use*” has the GP **V n as n**, and our goal is to derive the Chinese pattern such as ” 使用 $\mathbf{n_1}$ 作爲 $\mathbf{n_2}$” where ”$\mathbf{n_1}$” and ”$\mathbf{n_2}$” represent Chinese words corresponding to the first and second ”**n**” in **V n as n**.

The input to this stage is English-Chinese sentence pairs in a word-aligned bilingual parallel corpus, grammar pattern for each verb in each English sentence we extracted in the first stage (Step(1) in Figure 1), and translations of verbs in their GP we extracted in the third stage (Step(3) in Figure 1).

For each grammar pattern and verb, we extract Chinese counterparts of the English grammar pattern in all sentence pairs, and convert them into Chinese patterns. For example, the GP **use n as n** in the sentence ”*we use computer as a tool*”, corresponds to ” 使用電腦作爲工具” according to the word alignment, where the headword ”**use**” corresponds to ” 使用”(shih yong), the first ”**n**” corresponds to ” 電腦”(dian nao), ”**as**” corresponds to ” 作爲”(zuo wei), and the second ”**n**” corresponds to ” 工具”(gong jyu). ” 使用” and ” 作爲” which correspond to ”**V**” and ”**as**” and convert ” 電腦” and ” 工作” which correspond to the first ”**n**” and the second ”**n**” to ”$\mathbf{n_1}$” and ”$\mathbf{n_2}$” respectively . Therefore, Chinese pattern ” 使用 $\mathbf{n_1}$ 作爲 $\mathbf{n_2}$” is generated.

After generating Chinese patterns for each sentence, we count the number of each Chinese pattern for each English headword and GP as shown in Table 5, and calculate the average and standard deviation. We filter and identify Chinese patterns more frequent than average by 1 standard deviation.

The output of this stage is Chinese patterns for each English headword and GP.

4 Run-Time Interactive System

The system *Linggle GPTrans* is build on the foundation of *Linggle*, an linguistic search en-

Word	Trans	GP	Count	Std	Example
keep	跟上	**V n**	24	**3.32**	keep pace (跟上步伐)
keep	記錄	**V n**	52	**3.31**	keep track of expenses (記錄支出)
keep	記錄	V -ing	1	-0.24	-
keep	飼養	**V n**	26	**3.08**	keep animal (飼養動物)
keep	備存	**V n**	316	**3.29**	keep record (備存記錄)
keep	備存	V n -ing	7	-0.27	-
keep	備存	V n -ed	4	-0.29	-

Table 3: Count the number of each grammar pattern for each translation of “*keep*”

Word	Grammar pattern	Translations (Sorted by count)
charge	V n	收取 (212), 收費 (182), 徵收 (67), 指控 (20), 費用 (16), 起訴 (11), 檢控 (5), 落案 (2)
charge	V n n	收取 (29), 收費 (10)
charge	V to-inf	收費 (10), 收取 (4)
charge	V n with n	控告 (30), 指控 (11)
charge	V n for n	收取 (25), 收費 (14)
charge	V n to n	收取 (3)
keep	V n	保持 (1669), 維持 (356), 備存 (316), 保留 (241), 繼續 (147), 儲存 (121), 保存 (78), 保管 (55), 保住 (54), 記錄 (52), 留 (44), 留住 (40), 一直 (36), 保守 (33), 飼養 (26), 跟上 (24), 存備 (24), 阻止 (22), 持續 (13)
keep	V n form n	防止 (50), 阻止 (28), 保持 (16)
keep	V -ing	繼續 (233), 不斷 (163), 一直 (75), 不停 (59), 持續 (17), 下去 (13)
keep	V n -ed	保持 (31)
keep	V form n	阻止 (2), 保持 (2)
keep	V n -ing	保持 (42), 繼續 (19), 維持 (15)
keep	V to n	保持 (24), 維持 (13)
keep	V n to n	維持 (9), 保持 (8)
keep	V n as n	維持 (8), 保持 (7)

Table 4: Translations for each grammar pattern for “*charge*” and ”*keep*”

Word	Trans	GP	Chinese Pattern	Count
use	使用	V n as n	使用 n_1 作爲 n_2	35
use	使用	V n as n	使用 n_1 作 n_2	5
use	使用	V n as n	n_1 作爲 n_2 使用	3
use	使用	V n as n	使用 n_1	3
use	使用	V n as n	n_1 使用	1

Table 5: The Chinese patterns for the grammar pattern **use n as n** when translated as ”使用”

gine. *Linggle* indexes and retrieves common phrases in *Google Web 1T*. In addition to phrases, *Linggle* also provides example sentences extracted from *Google Books*.

We transform GPs to corresponding *Linggle* queries according to a phrase table (e.g. the notation **to-inf** is adjusted to **to v.**, and **pron-refl** is adjusted to **pron.** to fit our system). Additional adjustments are also made to make the patterns compatible to our system. Specifically, noun phrases in GP are translated into queries that retrieve pronouns, determiners, and nouns with leading adjectives. The symbol **wh** is transformed into question words (i.e. where, when, how, etc.). Specific verb tenses are identified by matching the corresponding suffix (i.e. -ing, -ed, etc.). For simplicity, we ignore light verb in our system since their meaning are heavily influenced by their objects. Table 6 shows sample queries for various GPs.

Generally, each GP of the head word has many translations, and some translations may have similar meanings. To avoid redundancy, we perform pairwise word vector similarity of the Chinese translations. Among the pairwise similarities of the Chinese translations, if a subset of translations have the similarity above 0.5, we drop the translation with lower ranking according to Section 3.3. To keep the system interface simple and information concise, we only display the top three translations for each GP.

We show the screenshot of the our system in Figure 4. The user input the word "keep" and the system presents the many GPs and relevant translations. One search result, "keep n from n" and their example sentences, are shown in Figure 4. We also provide a video demo of our system online.[2]

5 Evaluation

The purpose of *GPTrans* system is to allow users to retrieve the translation of GPs for the better understanding of a target pattern.The system automatically returns the translations for each GP of verbs to the users. Therefore, in this section, we report the results of preliminary evaluations on the extraction of GPs and their corresponding translations. The evaluation process was conducted on a set of verbs along with their GPs and translations of the GP.

[2] https://youtu.be/PQ-u07A5qM8

Figure 4: GPTrans search results for the pattern "keep n from n"

5.1 Experimental setting

The bilingual parallel corpora we used are the *Minutes of Legislative Council of the Hong Kong Special Administrative* from the legislative council of Hong Kong, and the *UM-corpus* from university of Macau. We used CKIP(Ma and Chen, 2003) which is a Chinese knowledge and information processing system developed by academic sinica to process Chinese word segmentation and used fast-align(Dyer et al., 2013) to process word alignment of bilingual parallel sentences.

We used Spacy(Honnibal and Montani, 2017) to parse English sentences and generated GPs for each verb in sentences as we described in section 3.1. Then, we extracted Chinese translations of each English word as we describe in section 3.2. Finally, we count and filter translations for each GP of each verb as we describe in section 3.3.

5.2 Evaluation Metrics

The output of our method are translations for each GP of all verbs in Collins dictionary. To evaluate our approach, we randomly selected 16 verbs with totally 126 GPs (388 items). The translations of each GP are evaluated by two linguists. Note that the verb *be* and light verbs were excluded from our evaluation since their senses usually depend on collocates. Among the selected sentence pairs, some English sentences are incorrectly parsed by SpaCy. That is, the target pattern of the verb does not exist

GP Tag	Linggle query
n	n./det._?adj._n./pron./pron._n.
wh	where/when/how/which/why/what
to-inf	to v.
-ing	$ing
-ed	$ed

Table 6: Partial phrase table for translation GP to Linggle queries

Word	Accuracy	Word	Accuracy
ask	93%	feel	90%
know	93%	make	81%
look	97%	train	88%
end	80%	agree	84%
argue	100%	answer	100%
save	96%	deal	66%
predict	100%	hang	100%
figure	0%	elect	88%

Table 7: Accuracy of each verbs.

in the English sentence. In such cases, these sentences pairs were removed from our evaluation set, resulting in the removal of 28 items and leaving 360 items in our evaluation. Then, we evaluated the average accuracy of translating each verb (shown in Table 7) and the overall accuracy. The overall accuracy is 91%.

5.3 Discussion

The result of evaluation shows that most of the translations generated by our method are correct. We observes that incorrect translations may be due to the incorrect parsing and incorrect part-of-speech (POS) tagging. Incorrect parsing may arises from tokenization of Chinese text which further leads to misalignment. POS errors occurs when one word can be multiple POS. For example, ”walk” could be a verb that denotes ”to move on foot” or a noun meanings ”a journey one make by walking.” Since grammar patterns are essentially verb phrases and noun phrases, incorrect POS tag would lead to erroneous results.

In addition, some training sentences are in passive voice, which lead to the detection of incorrect GPs. The GP in *COBUILD* are generally in active voice. The lack of passive voice GPs in predetermined rule patterns lead to detection of incorrect GPs.

Finally, GPs with consecutive noun phrases such as **V n n** are much more complicated to deal with since the two noun phrases are hard to distinguish from one another.

Looking at the results of our evaluation, the 0% accuracy for ”figure” stands out as an anomaly. After going through the testing data, we found out there are two reasons for the result. First of all, *COBUILD* treats the most common GP for ”figure”, **figure out n.** as a phrasal verb. Our system is build on the assumption that the verb would be a single word, thus dropping phrasal verbs in the process. Other than phrasal verbs, another reason for the result is incorrect POS tagging. Our system marks some occurrences of ”figure” as verbs when they are nouns in reality.

6 Conclusion and Future Work

Many avenues exist for future research and improvement of our system. We identify our system’s inability to process phrasal verbs during evaluation and plan to rectify this issue. One possible extension of our work is to utilize word embedding to consolidate translations with similar meanings but different wordings as one translation. We are also interested in extending our research outside of verbal grammar patterns and evacuating whether similar method would be effective for adjectives and nouns.

In summary, we discussed a method to provide translation of English grammar patterns for ESL learners and implemented a search system that shows the resulting translation. Our result presents nuanced translations for different GPs of the same word and helps ESL learners avoid common preposition errors. Our system utilized predetermined rule patterns and simple statistical models to identify the correct translation for each GP. This approach not only negates some of the reliance machine training models on training data size but also provides accurate translation results.

References

Joanne Boisson, Ting-Hui Kao, Jian-Cheng Wu, Tzu-Hsi Yen, and Jason S Chang. 2013. Linggle: a web-scale linguistic search engine for words in context. In *Proceedings of the 51st Annual Meeting of the Association for Computational*

Linguistics: System Demonstrations, pages 139–144.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Roberta Catizone, Graham Russell, and Susan Warwick. 1989. Deriving translation data from bilingual texts. In *Proceedings of the First International Lexical Acquisition Workshop*, pages 1–7.

Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Yi-Jyun Chen, Ching-Yu Helen Yang, and Jason S. Chang. 2020. Improve word alignment for extraction phrasal translations. *International Journal of Computational Linguistics Chinese Language Processing*, 25(2):37–54.

Collins Cobuild et al. 2005. *Collins Cobuild English Grammar*. Collins Cobuild.

Fathi Debili and Elyès Sammouda. 1992. Aligning sentences in bilingual texts french-english and french-arabic. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *arXiv preprint cmp-lg/9505016*.

Pascale Fung and Kenneth Church. 1994. K-vec: A new approach for aligning parallel texts. *arXiv preprint cmp-lg/9407021*.

William A Gale and Kenneth Church. 1991. Identifying word correspondences in parallel texts. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

William A Gale and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language acquisition*, 1(1):3–55.

Kathy Hirsh-Pasek and Roberta M Golinkoff. 1999. *The origins of grammar: Evidence from early language comprehension*. MIT press.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Mutsumi Imai, Lianjing Li, Etsuko Haryu, Hiroyuki Okada, Kathy Hirsh-Pasek, Roberta Michnick Golinkoff, and Jun Shigematsu. 2008. Novel noun and verb learning in chinese-, english-, and japanese-speaking children. *Child development*, 79(4):979–1000.

Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. *Computational linguistics*, 19(1):121–142.

M. H. Ko. 2006. Alignment of Multi-word Expressions in Parallel Corpora. Master's thesis, National Tsing Hua University, Taiwan.

Judit Kormos and Kata Csizer. 2014. The interaction of motivation, self-regulatory strategies, and autonomous learning behavior in different learner groups. *Tesol Quarterly*, 48(2):275–299.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to ckip chinese word segmentation system for the first international chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 168–171.

I Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. *arXiv preprint cmp-lg/9505044*.

I Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Robert C Moore. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Michel Simard, George F Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pages 1071–1082.

Sandra R Waxman and Amy E Booth. 2001. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. *Cognitive psychology*, 43(3):217–242.

Dekai Wu and Xuanyin Xia. 1994. Learning an english-chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*.

David Yarowsky. 1993. One sense per collocation. Technical report, PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE.